Statistical Methods for Large Complex Datasets

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Abhirup Datta

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Sudipto Banerjee, PhD and Hui Zou, PhD

May, 2016

Acknowledgments

My PhD dissertation would not have been sufficient and complete without the contribution of several individuals (no statistics pun intended). Dr. Sudipto Banerjee has been an amazing mentor for me. He gave me a lot of independence to develop my own projects and yet was always there to guide me in the right direction. Working with him has been a great learning experience and has played a huge role in my development as a researcher in spatial statistics. He also helped me improve my academic writing and presentation skills – often neglected aspects of research, that are extremely important. Sudipto funded me through research assistantships in the first six semesters of my PhD. Yet he offered me the flexibility to work on other projects. Sudipto has truly embodied the phrase "friend, philosopher and guide" for me during my PhD. Research meetings with him often metamorphosed into enthusiastic discussions about politics, passionate arguments on sports or simply nostalgic ramblings about our common alma mater. He has always looked out for me and never failed to inform me about any potential opportunity to progress my career. His optimism and guidance has helped me overcome many roadblocks during my PhD. I really hope that working with Sudipto has infused some of his research prowess and mentoring qualities in me which will be extremely valuable for my future career in academia.

I was also very fortunate to have Dr. Hui Zou as my co-advisor. I started working with Hui from my second year and under his guidance I was able to delve into research on statistical modeling of high-dimensional data. It was a new topic for me and was different from my ongoing research on spatial statistics. So the progress was often slow. Hui was really patient and has always helped me as I struggled managing different research projects. His expertise on the topic was absolutely critical for my development. I really enjoyed the weekly stimulating discussion sessions with him as well as working

out technical details together. Working with Hui has been a great privilege.

Dr. James S. Hodges has also played a big role in my PhD. Apart from being the best course instructor in my PhD, Jim has been a great mentor, research collaborator and is serving in my dissertation committee. Despite not being my advisor, Jim was always willing to meet me whenever I have requested and has provided marvelous insights even on my other research projects where he is not a co-author. He approached every problem from a very fundamental perspective and asked important and difficult research questions which forced me to think and learn more about the subject.

Dr. Andrew O. Finley, of Michigan State University, has been a long term collaborator and co-author in many of my dissertation projects. Andy's knowledge in the fields of forestry and geography and his expertise in large scale computations has considerably expedited the projects. I thank him for his invaluable role in my dissertation. I would also like to express my gratitude to Dr. Eric Lock for serving in my dissertation committee. Eric had also been a part of my preliminary oral examination committee and has lent thoughtful insights on my dissertation. Dr. Saonli Basu is my PhD academic advisor and has provided valuable guidance throughout my PhD. Saonli has also been a great host for me and many Bengali students in Minneapolis. Gatherings at her home with friends and Bengali food was often the high point amidst many stressful stretches during the PhD.

I am grateful to have the opportunity to pursue a career in research and academia after I graduate. This would not have been possible without the amazing help I received during my job application process. Sudipto, Hui, Jim and Andy have all kindly written several recommendation letters, sat through practice presentations and provided extremely valuable feedback and suggestions. I would also like to thank the Biostat Seminar Committee for arranging a practice job talk before my interviews. All these played a major role during my job applications and helped me get a job. In my final year of my PhD I was awarded the Interdisciplinary Fellowship Award from the Graduate School at University of Minnesota. I would like to thank the Institute on the Environment for endorsing me as an applicant and providing me with an office and conference travel funds. I would also like to thank Dr. Arindam Banerjee for kindly agreeing to be my mentor for the Fellowship.

Finally, without the support of my family it would have been impossible to follow

my dream and pursue a PhD. I thank my parents Mrs. Shila Datta and Mr. Adhikram Datta for the strength and encouragement they have given me. I would also like to express gratitude to my parents-in-law Mr. Pranab Ray and Mrs. Tapati Ray for their unfailing emotional support. My wife Dr. Debashree Ray has been always there to share the roller-coaster ride of PhD and life. She has been an endless source of strength for me over the last seven years and has inexplicably put up with all the ups and downs of my PhD while pursuing her own PhD and post-doctoral research. She keeps amazing me with her patience, belief and willingness to proof-read innumerable pages of my research.

Dedicated to my mother Mrs. Shila Datta and my loving wife Dr. Debashree Ray

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern technological advancements have enabled massive-scale collection, processing and storage of information, triggering the onset of the 'big data' era where in every two days now we create as much data as we did in the entire 20^{th} century. The information explosion has resulted in large and complex datasets which can potentially be exploited to seek solutions for relevant research problems. This, in turn, has necessitated the exponential growth of literature devoted to statistical modeling of 'big data'. For statisticians, big data is a relative all-encompassing term referring to datasets whose dimensions stretch the comfort levels of traditional statistical machinery. The theoretical and computational challenges posed by big data are of diverse nature and usually depend on the nature of the dataset and the associated statistical question. This thesis aims at developing novel statistical methods that can efficiently analyze a variety of such large complex datasets.

Underlying the umbrella theme of big data modeling, we present statistical methods for two different classes of large complex datasets. The first half of the thesis focuses on the 'large n' problem for large spatial or spatio-temporal datasets where observations exhibit strong dependencies across space and time. With the growing capabilities of Geographical Information Systems (GIS) and user-friendly software, statisticians today routinely encounter massive geographically referenced datasets from diverse areas of research like forestry, environmental health, climate sciences, finance, etc. containing a large number of irregularly located observations on multiple variables. Classical regression approaches based on the assumption of independent observations fail to capture the

space-varying dependence among the observations and result in inaccurate inference for such datasets. This has fueled considerable interest in statistical modeling for location-referenced spatial data. Gaussian Process (GP) models offer a rich modeling framework and are being widely deployed to help researchers comprehend complex spatial phenomena in the sciences. However, for large spatial datasets, the computational demands of traditional GP models are impossible to meet due to large matrix computations. Popular computationally attractive alternatives have been developed in recent times but many of them suffer from inferential limitations while others lack the versatility offered by process-based modeling.

In Chapter 2, we develop a Nearest Neighbor Gaussian Process (NNGP) that attains the balance between computational efficiency and inferential accuracy for massive spatial datasets. NNGP is constructed using the central idea that spatial dependence is stronger among neighboring locations. We establish that the NNGP is a well-defined spatial process providing legitimate finite-dimensional Gaussian densities with sparse precision matrices. We embed the NNGP as a sparsity-inducing prior within a rich hierarchical modeling framework and outline how computationally efficient Markov chain Monte Carlo (MCMC) algorithms can be executed without storing or decomposing large matrices. The floating point operations (flops) per iteration of this algorithm is linear in the number of spatial locations. Hence the model's scalability to massive datasets such as those found in climate sciences far exceeds those of process-based low rank models. More importantly, we have demonstrated that the NNGP does not oversmooth like low-rank models and effectively reproduces the corresponding inference from full (but highly expensive) GP models. We have also analyzed a massive forest biomass dataset observed at more than $100,000$ locations using the NNGP model. The results of our analysis lead to improved prediction of forest biomass based on satellite image data.

In modeling ambient air pollution data, it is now customary to meld observed measurements with physical model outputs, where the latter can operate at much finer scales. GP models are also commonly used to analyze such spatio-temporal data where inference is sought at arbitrary spatial and temporal resolutions. The lack of scalability of spatial GP models is inherited by their spatio-temporal analogs. In Chapter 3, we extend the work in Chapter 2 for large spatio-temporal datasets. One needs to account

for local dependence in both space and time for modeling such geographically and temporally referenced datasets. However, there is no unique concept of distance or local neighborhoods in a spatio-temporal domain. We construct dynamic local neighborhoods in a continuous spatio-temporal domain using strength of a correlation function as a proxy for distance. We develop a Dynamic Nearest Neighbor Gaussian Process (DNNGP) which enjoys sparse characterizations similar to the NNGP. DNNGP uses a novel scalable algorithm to simultaneously learn about the neighborhood structure along with the process parameters. Like the NNGP, DNNGP also scales linearly with the size of the dataset and delivers process-based inference at arbitrary resolutions for massive spatio-temporal datasets.

We use DNNGP model to create maps of particulate matter (PM), a class of malicious environmental pollutants known to cause detrimental effects on human health. Regulatory efforts aimed at curbing PM levels in different countries require high resolution space-time maps that can identify red-flag regions exceeding statutory concentration limits. Chemistry Transport Models (CTM) used to generate such maps have been shown to systematically underestimate observed PM concentrations. We propose a hierarchical DNNGP model that uses the CTM output to significantly improve prediction of PM levels across Europe. Additionally, the inference provided for the model covariance parameters provides insight into long-term space-time structures of PM.

Epidemiological data for different disease rates are often recorded as aggregated disease counts over entire geographical regions instead of isolated locations. Accurate identification of trends and factors associated with the disease requires accounting for the spatial dependence among the different regions in these areal datasets. The class of models for non-replicate areal data is currently very limited. Popularly used Conditional Autoregressive (CAR) models correspond to improper probability distributions and often lead to widely documented oversmoothed fits. In Chapter 4, we develop a new class of models based on directed acyclic graphs constructed over the geographical region. Unlike CAR models, our proposed Directed Acyclic Graph Autoregressive (DAGAR) models produce proper probability distributions without introducing any additional parameters. DAGAR models are also free of the ordering of the areal regions, unlike other approaches based on directed acyclic graph. DAGAR yields multivariate Gaussian distributions with sparse precision matrices and, hence, can be easily used

to model very large areal datasets. Extensive simulation experiments reveal that when the spatial signal is strong or when the latent surface is rough, DAGAR significantly outperforms CAR models.

In numerous application domains, the data is of considerably higher dimension and is not necessarily spatial. Researchers from diverse fields such as genetics, economics, neuroscience, public health, imaging, etc., are increasingly encountering complex datasets where dimension of each observation (p) substantially exceeds the size of the dataset (n). In the second half of this thesis we focus on regression problems in this "large p small n" setting. One important objective of high dimensional regression is to segregate a small set of regressors, associated with the response of interest, from the large number of redundant ones. Over the last two decades there has been a deluge of statistical methods aimed at accurately recovering this small set of associations in such high dimensional regression problems. However, the vast majority of such high-dimensional regression methods ignores possible contamination and heterogeneity of the data. We have developed statistical methods that are tailored to accommodate such aberrations.

We often face corrupted data in many applications where missing data and measurement errors cannot be ignored. For instance, microarrays containing information about thousands of genes are often contaminated with substantial error accruing because of the complexity of the data collection process. The Lasso is a popular variable selection method for clean data. The virtues of convexity have contributed fundamentally to the success and popularity of the Lasso as it is easily implemented using very fast solvers like the co-ordinate descent or homotopy algorithms. However, it has been proven that standard high dimensional regression techniques like Lasso may yield incorrect estimates when applied naively to such noisy datasets. In Chapter 5, we propose a new method named CoCoLasso (Convex Conditioned Lasso) that is convex and can handle a general class of corrupted datasets including the cases of additive measurement error and missing completely at random missing data. CoCoLasso adjusts for the contamination in a high-dimensional regression setup using convex optimization. We provide finite sample and asymptotic theoretical guarantees about parameter estimation and sparsity recovery. We also develop a simple algorithm using Alternating Direction Method of Multipliers (ADMM) to practically implement the method and devise novel cross validation and model comparison methods adapted to noisy datasets. We back our theory

with results using simulated datasets.

The US stock market crash in the 'Internet bubble burst' of 2001-2002 was not accompanied by plummeting house prices whereas in the sub-prime mortgage crisis in 2007-2009, stocks and house prices witnessed simultaneous collapse. Economic time series data are often collected over such heterogeneous regimes that are more appropriately modeled using piece-wise linear models for each segment of the data separated by change-points. Existing statistical literature on high dimensional regression often assumes a homogeneous model for the entire dataset. Change point methods for high dimensional regression data are severely underdeveloped. In addition to detecting the number and location of the change points, there is also interest in understanding the change in the association between the response and the predictors before and after a change point. In Chapter 6, we have developed a fully Bayesian framework for change-point high-dimensional linear regression where the slope of the regression can change after each change point. Using segment-specific shrinkage and diffusion priors, we deliver full posterior inference for the change points and simultaneously obtain posterior probabilities of variable selection in each segment via an efficient Gibbs sampler. Additionally, our method can detect an unknown number of change points and accommodate different variable selection constraints among the predictors like grouping or partial selection. We apply our approach for a macro-economic analysis of Minnesota house price index data. The results strongly favor the change point model over a homogeneous (no change point) high-dimensional regression model. Applicability of the approach extends beyond economic time-series to any high-dimensional dataset exhibiting heterogeneous trends like DNA micro-arrays and climate change data.

Chapter 2

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

2.1 Introduction

With the growing capabilities of Geographical Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced datasets containing a large number of irregularly located observations on multiple variables. This has, in turn, fueled considerable interest in statistical modeling for location-referenced spatial data; see, for example, the books by Stein (1999), Banerjee et al. (2014), Schabenberger and Gotway (2004), and Cressie and Wikle (2011) for a variety of methods and applications. Spatial process models introduce spatial dependence between observations using an underlying random field, $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, over a region of interest $\mathcal{D}$, which is endowed with a probability law that specifies the joint distribution for any finite set of random variables. For example, a zero-centered Gaussian process ensures that $\mathbf{w} = (w(\mathbf{s}_1), w(\mathbf{s}_2) \ldots, w(\mathbf{s}_n))' \sim N(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$, where $\mathbf{C}(\boldsymbol{\theta})$ is a family of covariance matrices, indexed by an unknown set of parameters $\boldsymbol{\theta}$. Such processes offer a rich modeling framework and are being widely deployed to help researchers comprehend complex spatial phenomena in the sciences. However, model fitting usually involves the

inverse and determinant of $\mathbf{C}(\boldsymbol{\theta})$, which typically require $\sim n^3$ floating point operations (flops) and storage of the order of n^2. These become prohibitive when n is large and $\mathbf{C}(\boldsymbol{\theta})$ has no exploitable structure.

Broadly speaking, modeling large spatial datasets proceeds from either exploiting "low-rank" models or using sparsity. The former attempts to construct spatial processes on a lower-dimensional subspace (see, e.g., Higdon, 2001; Kammann and Wand, 2003; Stein, 2007, 2008; Banerjee et al., 2008; Cressie and Johannesson, 2008; Crainiceanu et al., 2008; Rasmussen and Williams, 2005; Finley et al., 2009) by regressing the original (*parent*) process on its realizations over a smaller set of $r << n$ locations ("knots" or "centers"). The algorithmic cost for model fitting typically decreases from $O(n^3)$ to $O(nr^2 + r^3) \approx O(nr^2)$ flops since $n >> r$. However, when n is large, empirical investigations suggest that r must be fairly large to adequately approximate the parent process and the nr^2 flops becomes exorbitant (see Section 2.5.1). Furthermore, low rank models perform poorly when neighboring observations are strongly correlated and the spatial signal dominates the noise (Stein, 2014). Although bias-adjusted low-rank models tend to perform better (Finley et al., 2009; Banerjee et al., 2010; Sang and Huang, 2012), they increase the computational burden.

Sparse methods include covariance tapering (see, e.g., Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009; Shaby and Ruppert, 2012), which introduces sparsity in $\mathbf{C}(\boldsymbol{\theta})$ using compactly supported covariance functions. This is effective for parameter estimation and interpolation of the response ("kriging"), but it has not been fully developed or explored for more general inference on residual or latent processes. Introducing sparsity in $\mathbf{C}(\boldsymbol{\theta})^{-1}$ is prevalent in approximating Gaussian process likelihoods using Markov random fields (e.g., Rue and Held, 2005), products of lower dimensional conditional distributions (Vecchia, 1988, 1992; Stein et al., 2004), or composite likelihoods (e.g., Bevilacqua and Gaetan, 2014; Eidsvik et al., 2014). However, unlike low rank processes, these do not, necessarily, extend to new random variables at arbitrary locations. There may not be a corresponding process, which restricts inference to the estimation of spatial covariance parameters. Spatial prediction ("kriging") at arbitrary locations proceeds by imputing estimates into an interpolator derived from a different process model. This may not reflect accurate estimates of predictive uncertainty and is undesirable.

Our intended inferential contribution is to offer substantial scalability for fully process-based inference on underlying, perhaps completely unobserved, spatial processes. Moving from finite-dimensional sparse likelihoods to sparsity-inducing spatial processes can be complicated. We first introduce sparsity in finite-dimensional probability models using specified neighbor sets constructed from directed acyclic graphs. We use these sets to extend these finite-dimensional models to a valid spatial process over uncountable sets. We call this process a *Nearest-Neighbor Gaussian Process* (NNGP). Its finite-dimensional realizations have sparse precision matrices available in closed form. While sparsity has been effectively exploited by Vecchia (1988); Stein et al. (2004); Emory (2009); Gramacy and Apley (2014); Gramacy et al. (2014) and Stroud et al. (2014) for approximating expensive likelihoods cheaply, a fully process-based modeling and inferential framework has, hitherto, proven elusive. The NNGP fills this gap and enriches the inferential capabilities of existing methods by subsuming estimation of model parameters, prediction of outcomes and interpolation of underlying processes into one highly scalable unifying framework.

To demonstrate its full inferential capabilities, we deploy the NNGP as a sparsity-inducing prior for spatial processes in a Bayesian framework. Unlike low rank processes, the NNGP always specifies non-degenerate finite dimensional distributions making it a legitimate proper prior for random fields and is applicable to any class of distributions that support a spatial stochastic process. It can, therefore, model an underlying process that is never actually observed. The modeling provides structured dependence for random effects, e.g. intercepts or coefficients, at a second stage of specification where the first stage need not be Gaussian. We cast a multivariate NNGP within a versatile spatially-varying regression framework (Gelfand et al., 2003; Banerjee et al., 2008) and conveniently obtain entire posteriors for all model parameters as well as for the spatial processes at both observed and unobserved locations. Using a forestry example, we show how the NNGP delivers process-based inference for spatially-varying regression models at a scale where even low-rank processes, let alone full Gaussian processes, are unimplementable even in high-performance computing environments.

Here is a brief outline. Section 2.2 formulates the NNGP using multivariate Gaussian processes. Section 2.3 outlines Bayesian estimation and prediction within a very flexible hierarchical modeling setup. Section 2.4 discusses alternative NNGP models and

algorithms. Section 2.5 presents simulation studies to highlight the inferential benefits of the NNGP and also analyzes forest biomass from a massive USDA dataset. Finally, Section 2.6 concludes the manuscript with a brief summary and pointers toward future work.

2.2 Nearest-Neighbor Gaussian Process

2.2.1 Gaussian density on sparse directed acyclic graphs

We will consider a q-variate spatial process over $\Re^d$. Let $\mathbf{w}(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$ denote a zero-centered q-variate Gaussian process, where $\mathbf{w}(\mathbf{s}) \in \Re^q$ for all $\mathbf{s} \in \mathcal{D} \subseteq \Re^d$. The process is completely specified by a valid cross-covariance function $\mathbf{C}(\cdot,\cdot \,|\, \boldsymbol{\theta})$, which maps a pair of locations $\mathbf{s}$ and $\mathbf{t}$ in $\mathcal{D} \times \mathcal{D}$ into a $q \times q$ real valued matrix $\mathbf{C}(\mathbf{s},\mathbf{t})$ with entries $\text{cov}\{w_i(\mathbf{s}), w_j(\mathbf{t})\}$. Here, $\boldsymbol{\theta}$ denotes the parameters associated with the cross-covariance function. Let $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ be a fixed collection of distinct locations in $\mathcal{D}$, which we call the *reference set*. So, $\mathbf{w}_\mathcal{S} \sim N(\mathbf{0}, \mathbf{C}_\mathcal{S}(\boldsymbol{\theta}))$, where $\mathbf{w}_\mathcal{S} = (\mathbf{w}(\mathbf{s}_1)', \mathbf{w}(\mathbf{s}_2)', \ldots, \mathbf{w}(\mathbf{s}_k)')'$ and $\mathbf{C}_\mathcal{S}(\boldsymbol{\theta})$ is a positive definite $qk \times qk$ block matrix with $\mathbf{C}(\mathbf{s}_i, \mathbf{s}_j)$ as its blocks. Henceforth, we write $\mathbf{C}_\mathcal{S}(\boldsymbol{\theta})$ as $\mathbf{C}_\mathcal{S}$, the dependence on $\boldsymbol{\theta}$ being implicit, with similar notation for all spatial covariance matrices.

The reference set $\mathcal{S}$ need not coincide with or be a part of the observed locations, so k need not equal n, although we later show that the observed locations are a convenient practical choice for $\mathcal{S}$. When k is large, parameter estimation becomes computationally cumbersome, perhaps even unfeasible, because it entails the inverse and determinant of $\tilde{\mathbf{C}}_\mathcal{S}$. Here, we benefit from expressing the joint density of $\mathbf{w}_\mathcal{S}$ as the product of conditional densities, i.e.,

$$p(\mathbf{w}_\mathcal{S}) = p(\mathbf{w}(\mathbf{s}_1))\; p(\mathbf{w}(\mathbf{s}_2) \,|\, \mathbf{w}(\mathbf{s}_1))\; \ldots\; p(\mathbf{w}(\mathbf{s}_k) \,|\, \mathbf{w}(\mathbf{s}_{k-1}), \ldots, \mathbf{w}(\mathbf{s}_1))\;, \tag{2.1}$$

and replacing the larger conditioning sets on the right hand side of (2.1) with smaller, carefully chosen, conditioning sets of size at most m, where $m \ll k$ (see, e.g., Vecchia, 1988; Stein et al., 2004; Gramacy and Apley, 2014; Gramacy et al., 2014). So, for every $\mathbf{s}_i \in \mathcal{S}$, a smaller conditioning set $N(\mathbf{s}_i) \subset \mathcal{S} \setminus \{\mathbf{s}_i\}$ is used to construct

$$\tilde{p}(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{w}_{N(\mathbf{s}_i)})\;, \tag{2.2}$$

where $\mathbf{w}_{N(\mathbf{s}_i)}$ is the vector formed by stacking the realizations of $\mathbf{w}(\mathbf{s})$ over $N(\mathbf{s}_i)$.

Let $N_{\mathcal{S}} = \{N(\mathbf{s}_i); i = 1, 2, \ldots, k\}$ be the collection of all conditioning sets over $\mathcal{S}$. We can view the pair $\{\mathcal{S}, N_{\mathcal{S}}\}$ as a directed graph $\mathcal{G}$ with $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ being the set of nodes and $N_{\mathcal{S}}$ the set of directed edges. For every two nodes $\mathbf{s}_i$ and $\mathbf{s}_j$, we say $\mathbf{s}_j$ is a directed neighbor of $\mathbf{s}_i$ if there is a directed edge from $\mathbf{s}_i$ to $\mathbf{s}_j$. So, $N(\mathbf{s}_i)$ denotes the set of directed neighbors of $\mathbf{s}_i$ and is, henceforth, referred to as the "neighbor set" for $\mathbf{s}_i$. A "directed cycle" in a directed graph is a chain of nodes $\mathbf{s}_{i_1}, \mathbf{s}_{i_2}, \ldots, \mathbf{s}_{i_b}$ such that $\mathbf{s}_{i_1} = \mathbf{s}_{i_b}$ and there is a directed edge between $\mathbf{s}_{i_j}$ and $\mathbf{s}_{i_{j+1}}$ for every $j = 1, 2, \ldots, b-1$. A directed graph with no directed cycles is known as a 'directed acyclic graph'.

We now show that if $\mathcal{G} = (\mathcal{S}, N_{\mathcal{S}})$ is acyclic then $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ defined in (2.3) corresponds to a true density over $\mathcal{S}$. For any directed acyclic graph, there exists a node with zero in-degree i.e. no directed edge pointing towards it. We denote this node by $\mathbf{s}_{\pi(1)}$ This means $\mathbf{s}_{\pi(1)}$ does not belong to the neighbor set of any other location in $\mathcal{S}$. The only term where it appears on the right hand side of (2.2) is $p(\mathbf{w}(\mathbf{s}_{\pi(1)} \,|\, \mathbf{w}_{N(\mathbf{s}_{\pi(1)})})$ which integrates out to one with respect to $d\mathbf{w}(\mathbf{s}_{\pi(1)})$. We now have a new acyclic directed graph $\mathcal{G}_1$ obtained by removing vertex $\mathbf{s}_{\pi(1)}$ and its directed edges from $\mathcal{G}$. Now we can find a new vertex $\mathbf{s}_{\pi(2)}$ with zero out-degree in $\mathcal{G}_1$ and continue as before to get a permutation $\pi(1), \pi(2), \ldots, \pi(k)$ of $1, 2, \ldots, k$ such that

$$\int \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{w}_{N(\mathbf{s}_i)}) d\mathbf{w}(\mathbf{s}_{\pi(1)}) d\mathbf{w}(\mathbf{s}_{\pi(2)}) \ldots d\mathbf{w}(\mathbf{s}_{\pi(k)}) = 1$$

An easy application of Fubini's theorem now ensures that this is a proper joint density.

Hence, if $\mathcal{G}$ is a directed acyclic graph, then $\tilde{p}(\mathbf{w}_{\mathcal{S}})$, as defined above, is a proper multivariate joint density (see Lauritzen (1996) for a similar result). Starting from a joint multivariate density $p(\mathbf{w}_{\mathcal{S}})$, we derive a new density $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ using a directed acyclic graph $\mathcal{G}$. While this holds for any original density $p(\mathbf{w}_{\mathcal{S}})$, it is especially useful in our context, where $p(\mathbf{w}_{\mathcal{S}})$ is a multivariate Gaussian density and $\mathcal{G}$ is sufficiently sparse. To be precise, let $\mathbf{C}_{N(\mathbf{s}_i)}$ be the covariance matrix of $\mathbf{w}_{N(\mathbf{s}_i)}$ and let $\mathbf{C}_{\mathbf{s}_i, N(\mathbf{s}_i)}$ be the $q \times mq$ cross-covariance matrix between the random vectors $\mathbf{w}(\mathbf{s}_i)$ and $\mathbf{w}_{N(\mathbf{s}_i)}$. Standard distribution theory reveals

$$\tilde{p}(\mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{k} N(\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{B}_{\mathbf{s}_i} \mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i}) \,, \tag{2.3}$$

where $\mathbf{B}_{\mathbf{s}_i} = \mathbf{C}_{\mathbf{s}_i,N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}$ and $\mathbf{F}_{\mathbf{s}_i} = \mathbf{C}(\mathbf{s}_i,\mathbf{s}_i) - \mathbf{C}_{\mathbf{s}_i,N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}\mathbf{C}_{N(\mathbf{s}_i),\mathbf{s}_i}$. So, the likelihood in (2.2) is proportional to

$$\frac{1}{\prod_{i=1}^{k}\sqrt{\det(\mathbf{F}_{\mathbf{s}_i})}}\exp\left(-\frac{1}{2}\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}^{-1}_{\mathbf{s}_i}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})\right)$$

For any matrix $\mathbf{A}$, let $\mathbf{A}[,j:j']$ denote the submatrix formed using columns j to j' where $j<j'$. For $j=1,2,\ldots,k$, we define $q\times q$ blocks $\mathbf{B}_{\mathbf{s}_i,j}$ as

$$\mathbf{B}_{\mathbf{s}_i,j} = \begin{cases} \mathbf{I}_q \text{ if } j=i; \\ -\mathbf{B}_{\mathbf{s}_i}[,(l-1)q+1:lq] \text{ if } \mathbf{s}_j = N(\mathbf{s}_i)(l) \text{ for some } l; \\ \mathbf{O} \text{ otherwise,} \end{cases}$$

where, for any location $\mathbf{s}$, $N(\mathbf{s})(l)$ is the l-th neighbor of $\mathbf{s}$. So, $\mathbf{w}_{\mathbf{s}_i}-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)} = \mathbf{B}^*_{\mathbf{s}_i}\mathbf{w}_{\mathcal{S}}$, where $\mathbf{B}^*_{\mathbf{s}_i} = [\mathbf{B}_{\mathbf{s}_i,1},\mathbf{B}_{\mathbf{s}_i,2},\ldots,\mathbf{B}_{\mathbf{s}_i,k}]$ is $q\times kq$ and sparse with at most $m+1$ non-zero blocks. Then,

$$\begin{aligned}&\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}^{-1}_{\mathbf{s}_i}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}) \\ &= \sum_{i=1}^{k}\mathbf{w}'_{\mathcal{S}}(\mathbf{B}^*_{\mathbf{s}_i})'\mathbf{F}^{-1}_{\mathbf{s}_i}\mathbf{B}^*_{\mathbf{s}_i}\mathbf{w}_{\mathcal{S}} = \mathbf{w}'_{\mathcal{S}}\mathbf{B}'_{\mathcal{S}}\mathbf{F}^{-1}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}\mathbf{w}_{\mathcal{S}}\ ,\end{aligned}$$

where $\mathbf{F} = diag(\mathbf{F}_{\mathbf{s}_1},\mathbf{F}_{\mathbf{s}_2},\ldots,\mathbf{F}_{\mathbf{s}_k})$ and $\mathbf{B}_{\mathcal{S}} = ((\mathbf{B}^*_{\mathbf{s}_1})',(\mathbf{B}^*_{\mathbf{s}_2})',\ldots,(\mathbf{B}^*_{\mathbf{s}_k})')'$. So, defining

$$\tilde{\mathbf{C}}_{\mathcal{S}} = (\mathbf{B}'_{\mathcal{S}}\mathbf{F}^{-1}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}})^{-1} \tag{2.4}$$

we have shown that $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ in (2.3) is a multivariate Gaussian density with covariance matrix $\tilde{\mathbf{C}}_{\mathcal{S}}$, which, obviously, is different from $\mathbf{C}_{\mathcal{S}}$.

From the form of $\mathbf{B}_{\mathbf{s}_i,j}$, it is clear that $\mathbf{B}_{\mathcal{S}}$ is sparse and lower triangular with ones on the diagonals. So, $\det(\mathbf{B}_{\mathcal{S}})=1$, $\det((\mathbf{B}'_{\mathcal{S}}\mathbf{F}^{-1}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}})^{-1}) = \prod\det(\mathbf{F}_{\mathbf{s}_i})$. Let $\tilde{\mathbf{C}}^{ij}_{\mathcal{S}}$ denote the $(i,j)^{th}$ block of $\tilde{\mathbf{C}}^{-1}_{\mathcal{S}}$. Then from equation (2.4) we see that for $i<j$, $\tilde{\mathbf{C}}^{ij}_{\mathcal{S}} = \sum_{l=j}^{k}(\mathbf{B}^*_{\mathbf{s}_l,i})'\mathbf{F}^{-1}_{\mathbf{s}_l}\mathbf{B}^*_{\mathbf{s}_l,j}$. So, $\tilde{\mathbf{C}}^{ij}_{\mathcal{S}}$ is non-zero only if there exists at least one location $\mathbf{s}_l$ such that $\mathbf{s}_i\in N(\mathbf{s}_l)$ and $\mathbf{s}_j$ is either equal to $\mathbf{s}_l$ or is in $N(\mathbf{s}_l)$. Since every neighbor set has at most m elements where $m\ll k$, there are at most $km(m+1)/2$ such pairs (i,j). So $\tilde{\mathbf{C}}^{-1}_{\mathcal{S}}$ is sparse with at most $km(m+1)q^2/2$ non-zero entries. Thus, for a very general class of neighboring sets, $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ defined in (2.2) is the joint density of a multivariate Gaussian distribution with a sparse precision matrix.

Turning to the neighbor sets, choosing $N(\mathbf{s}_i)$ to be any subset of $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ ensures an acyclic $\mathcal{G}$ and, hence, a valid probability density in (2.3). Several special cases exist in likelihood approximation contexts. For example, Vecchia (1988) and Stroud et al. (2014) specified $N(\mathbf{s}_i)$ to be the m nearest neighbors of $\mathbf{s}_i$ among $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}$ with respect to Euclidean distance. Stein et al. (2004) considered nearest as well as farthest neighbors from $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$. Gramacy and Apley (2014) offer greater flexibility in choosing $N(\mathbf{s}_i)$, but may require several approximations to be efficient.

All of the above choices depend upon an ordering of the locations. Spatial locations are not ordered naturally, so one imposes order by, for example, ordering on one of the coordinates. Of course, any other function of the coordinates can be used to impose order. However, the aforementioned authors have cogently demonstrated that the choice of the ordering has no discernible impact on the approximation of (2.1) by (2.3). Our own simulation experiments (see Section 2.5.2) concur with these findings; inference based upon $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ is extremely robust to the ordering of the locations. This is not entirely surprising. Clearly, whatever order we choose in (2.1), $p(\mathbf{w}_{\mathcal{S}})$ produces the full joint density. Note that we reduce (2.1) to (2.2) based upon neighbor sets constructed with respect to the *specific* ordering in (2.1). A different ordering in (2.1) will produce a different set of neighbors for (2.2). Since $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ ultimately relies upon the information borrowed from the neighbors, its effectiveness is often determined by the number of neighbors we specify and *not* the specific ordering.

In the following section, we will extend the density $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ to a legitimate spatial process. We remark that our subsequent development holds true for any choice of $N(\mathbf{s}_i)$ that ensures an acyclic $\mathcal{G}$. In general, identifying a "best subset" of m locations for obtaining optimal predictions for $\mathbf{s}_i$ is a non-convex optimization problem, which is difficult to implement and defeats our purpose of using smaller conditioning sets to ease computations. Nevertheless, we have found Vecchia's choice of m-nearest neighbors from $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ to be simple and to perform extremely well for a wide range of simulation experiments. In what ensues, this will be our choice for $N(\mathbf{s}_i)$ and the corresponding density $\tilde{p}(\mathbf{w}_{\mathcal{S}})$ will be referred to as the 'nearest neighbor' density of $\mathbf{w}_{\mathcal{S}}$.

2.2.2 Extension to a Gaussian Process

Let $\mathbf{u}$ be any location in $\mathcal{D}$ outside $\mathcal{S}$. Consistent with the definition of $N(\mathbf{s}_i)$,

let $N(\mathbf{u})$ be the set of m-nearest neighbors of $\mathbf{u}$ in $\mathcal{S}$. Hence, for any finite set $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ such that $\mathcal{S} \cap \mathcal{U}$ is empty, we define the nearest neighbor density of $\mathbf{w}_\mathcal{U}$ conditional on $\mathbf{w}_\mathcal{S}$ as

$$\tilde{p}(\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S}) = \prod_{i=1}^{r} p(\mathbf{w}(\mathbf{u}_i) \,|\, \mathbf{w}_{N(\mathbf{u}_i)}) \,. \tag{2.5}$$

This conditional density is akin to (2.2) except that all the neighbor sets are subsets of $\mathcal{S}$. This ensures a proper conditional density. Indeed (2.2) and (2.5) are sufficient to describe the joint density of *any* finite set over the domain $\mathcal{D}$. More precisely, if $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is *any* finite subset in $\mathcal{D}$, then, using (2.5) we obtain the density of $\mathbf{w}_\mathcal{V}$ as,

$$\tilde{p}(\mathbf{w}_\mathcal{V}) = \int \tilde{p}(\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S})\, \tilde{p}(\mathbf{w}_\mathcal{S}) \prod_{\{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}\}} d(\mathbf{w}(\mathbf{s}_i)) \;\; \text{where } \mathcal{U} = \mathcal{V} \setminus \mathcal{S} \,. \tag{2.6}$$

If $\mathcal{U}$ is empty, then (2.5) implies that $\tilde{p}(\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S}) = 1$ in (2.6). If $\mathcal{S} \setminus \mathcal{V}$ is empty, then the integration in (2.6) is not needed.

We now prove that these probability densities, defined on finite topologies, conform to Kolmogorov's consistency criteria and, hence, correspond to a valid spatial process over $\mathcal{D}$. We will first show that for every finite set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ in $\mathcal{D}$, $n \in \{1, 2, \ldots\}$ and for every permutation $\pi(1), \pi(2), \ldots, \pi(n)$ of $1, 2, \ldots, n$ we have,

$$\tilde{p}\left(\mathbf{w}(\mathbf{v}_1), \mathbf{w}(\mathbf{v}_2), \ldots, \mathbf{w}(\mathbf{v}_n)\right) = \tilde{p}\left(\mathbf{w}(\mathbf{v}_{\pi(1)}), \mathbf{w}(\mathbf{v}_{\pi(2)}), \ldots, \mathbf{w}(\mathbf{v}_{\pi(n)})\right) \,.$$

We begin by showing that for any finite set $\mathcal{V}$, the expression given in (2.6) is a proper density. Let $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$. Since $\mathcal{V} \cup (\mathcal{S} \setminus \mathcal{V}) = \mathcal{S} \cup \mathcal{U}$, we obtain

$$\int \tilde{p}(\mathbf{w}_\mathcal{V}) \prod_{\mathbf{v}_i \in \mathcal{V}} d(\mathbf{w}(\mathbf{v}_i)) = \int \tilde{p}(\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S}) \tilde{p}(\mathbf{w}_\mathcal{S}) \prod_{\mathbf{v}_i \in \mathcal{U}} d(\mathbf{w}(\mathbf{v}_i)) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i))$$

$$= \int \tilde{p}(\mathbf{w}_\mathcal{S}) \left(\int \tilde{p}(\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S}) \prod_{\mathbf{v}_i \in \mathcal{U}} d(\mathbf{w}(\mathbf{v}_i)) \right) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i)) = \int \tilde{p}(\mathbf{w}_\mathcal{S}) \prod_{\mathbf{s}_i \in \mathcal{S}} d(\mathbf{w}(\mathbf{s}_i)) = 1$$

Note that $\mathcal{S}$ is fixed. Therefore, the expression for the joint density of $\mathbf{w}_\mathcal{V}$ depends only on the the neighbor sets $N(\mathbf{v}_i)$ for $\mathbf{v}_i \in \mathcal{U}$. So the NNGP density for $\mathcal{V}$ is invariant under any permutation of locations inside $\mathcal{V}$.

We now prove that for every location $\mathbf{v}_0 \in \mathcal{D}$, $\tilde{p}(\mathbf{w}_\mathcal{V}) = \int \tilde{p}(\mathbf{w}_{\mathcal{V} \cup \{\mathbf{v}_0\}}) d(\mathbf{w}(\mathbf{v}_0))$. Let $\mathcal{V}_1 = \mathcal{V} \cup \{\mathbf{v}_0\}$. We split the proof into two cases. If $\mathbf{v}_0 \in \mathcal{S}$, then using the fact

$\mathcal{V}_1 \setminus \mathcal{S} = \mathcal{V} \setminus \mathcal{S} = \mathcal{U}$, we obtain

$$\int \tilde{p}(\mathbf{w}_{\mathcal{V}_1}) d(\mathbf{w}(\mathbf{v}_0)) = \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \tilde{p}(\mathbf{w}_{\mathcal{V}_1 \setminus \mathcal{S}} \,|\, \mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}_1} d(\mathbf{w}(\mathbf{s}_i)) d(\mathbf{w}(\mathbf{v}_0)$$
$$= \int \tilde{p}(\mathbf{w}_{\mathcal{S}}) \tilde{p}(\mathbf{w}_{\mathcal{V} \setminus \mathcal{S}} \,|\, \mathbf{w}_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}} d(\mathbf{w}(\mathbf{s}_i)) = \tilde{p}(\mathbf{w}_{\mathcal{U}}) \; .$$

If $\mathbf{v}_0 \notin \mathcal{S}$, then $\mathbf{w}(\mathbf{v}_0)$ does not appear in the neighborhood set of any other term. So, $p(\mathbf{w}(\mathbf{v}_0) \,|\, \mathbf{w}_{\mathcal{S}})$ integrates to one with respect to $d(\mathbf{w}(\mathbf{v}_0))$. The result now follows from $\int p(\mathbf{w}_{\mathcal{V}_1} \,|\, \mathbf{w}_{\mathcal{S}}) d(\mathbf{w}(\mathbf{v}_0)) = p(\mathbf{w}_{\mathcal{V}} \,|\, \mathbf{w}_{\mathcal{S}})$.

So, given any original (parent) spatial process and any *fixed* reference set $\mathcal{S}$, we can construct a new process over the domain $\mathcal{D}$ using a collection of neighbor sets in $\mathcal{S}$. We refer to this process as the 'nearest neighbor process' derived from the original parent process. If the parent process is $GP(\mathbf{0}, \mathbf{C}(\cdot, \cdot \,|\, \boldsymbol{\theta}))$, then

$$\tilde{p}(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{r} N(\mathbf{w}(\mathbf{u}_i) \,|\, \mathbf{B}_{\mathbf{u}_i} \mathbf{w}_{N(\mathbf{u}_i)}, \mathbf{F}_{\mathbf{u}_i}) = N(\mathbf{B}_{\mathcal{U}} \mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}}) \tag{2.7}$$

for any finite set $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ in $\mathcal{D}$ outside $\mathcal{S}$, where $\mathbf{B}_{\mathbf{u}_i}$ and $\mathbf{F}_{\mathbf{u}_i}$ are defined analogous to (2.3) based on the neighbor sets $N(\mathbf{u}_i)$, $\mathbf{F}_{\mathcal{U}} = diag(\mathbf{F}_{\mathbf{u}_1}, \mathbf{F}_{\mathbf{u}_2}, \ldots, \mathbf{F}_{\mathbf{u}_r})$ and $\mathbf{B}_{\mathcal{U}}$ is a sparse $nq \times kq$ matrix with each row having at most mq non-zero entries. For any finite set $\mathcal{V}$ in $\mathcal{D}$, $\tilde{p}(\mathbf{w}_{\mathcal{V}})$ is the density of the realizations of a Gaussian Process over $\mathcal{V}$ with cross covariance function

$$\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2; \boldsymbol{\theta}) = \begin{cases} \tilde{\mathbf{C}}_{\mathbf{s}_i, \mathbf{s}_j} \; , & \text{if } \mathbf{v}_1 = \mathbf{s}_i \text{ and } \mathbf{v}_2 = \mathbf{s}_j \text{ are both in } \mathcal{S}, \\ \mathbf{B}_{\mathbf{v}_1} \tilde{\mathbf{C}}_{N(\mathbf{v}_2), \mathbf{s}_j} & \text{if } \mathbf{v}_1 \notin \mathcal{S} \text{ and } \mathbf{v}_2 = \mathbf{s}_j \in \mathcal{S}, \\ \mathbf{B}_{\mathbf{v}_1} \tilde{\mathbf{C}}_{N(\mathbf{v}_1), N(\mathbf{v}_2)} \mathbf{B}'_{\mathbf{v}_2} + \delta_{(\mathbf{v}_1 = \mathbf{v}_2)} \mathbf{F}_{\mathbf{v}_1} & \text{if } \mathbf{v}_1 \text{ and } \mathbf{v}_2 \text{ are not in } \mathcal{S} \end{cases} \tag{2.8}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are any two locations in $\mathcal{D}$, $\tilde{\mathbf{C}}_{A,B}$ denotes submatrices of $\tilde{\mathbf{C}}_{\mathcal{S}}$ indexed by the locations in the sets A and B, and $\delta_{(\mathbf{v}_1 = \mathbf{v}_2)}$ is the Kronecker delta.

This completes the construction of a well-defined *Nearest Neighbor Gaussian Process*, $NNGP(\mathbf{0}, \tilde{\mathbf{C}}(\cdot, \cdot \,|\, \boldsymbol{\theta}))$, derived from a *parent Gaussian process*, $GP(\mathbf{0}, \mathbf{C}(\cdot, \cdot \,|\, \boldsymbol{\theta}))$. In the NNGP, the size of $\mathcal{S}$, i.e., k, can be as large, or even larger than the size of the dataset. The reduction in computational complexity is achieved through sparsity of the NNGP precision matrices. Unlike low-rank processes, the NNGP is *not* a degenerate process. It is a proper, sparsity-inducing Gaussian process, immediately available as

a prior in hierarchical modeling, and, as we show in the next section, delivers massive computational benefits.

2.3 Bayesian estimation and implementation

2.3.1 A hierarchical model

Consider a vector of l dependent variables, say $\mathbf{y}(\mathbf{t})$, at location $\mathbf{t} \in \mathcal{D} \subseteq \Re^d$ in a spatially-varying regression model,

$$\mathbf{y}(\mathbf{t}) = \mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t}) + \boldsymbol{\epsilon}(\mathbf{t}) \,, \tag{2.9}$$

where $\mathbf{X}(\mathbf{t})'$ is the $l \times p$ matrix of fixed spatially-referenced predictors, $\mathbf{w}(\mathbf{t})$ is a $q \times 1$ spatial process forming the coefficients of the $l \times q$ fixed design matrix $\mathbf{Z}(\mathbf{t})'$, and $\boldsymbol{\epsilon}(\mathbf{t}) \overset{iid}{\sim} N(\mathbf{0}, \mathbf{D})$ is an $l \times 1$ white noise process capturing measurement error or micro-scale variability with dispersion matrix $\mathbf{D}$, which we assume is diagonal with entries τ_j^2, $j = 1, 2, \ldots, l$. The matrix $\mathbf{X}(\mathbf{t})'$ is block diagonal with $p = \sum_{i=1}^{l} p_i$, where the $1 \times p_i$ vector $\mathbf{x}_i(\mathbf{t})'$, including perhaps an intercept, is the i-th block for each $i = 1, 2, \ldots, l$. The model in (2.9) subsumes several specific spatial models. For instance, letting $q = l$ and $\mathbf{Z}(\mathbf{t})' = \mathbf{I}_{l \times l}$ leads to a multivariate spatial regression model where $\mathbf{w}(\mathbf{t})$ acts as a *spatially-varying intercept*. On the other hand, we could envision all coefficients to be spatially-varying and set $q = p$ with $\mathbf{Z}(\mathbf{t})' = \mathbf{X}(\mathbf{t})'$.

For scalability, instead of a customary Gaussian process prior for $\mathbf{w}(t)$ in (2.9), we assume $\mathbf{w}(\mathbf{t}) \sim NNGP(\mathbf{0}, \tilde{\mathbf{C}}(\cdot, \cdot \,|\, \boldsymbol{\theta}))$ derived from the parent $GP(\mathbf{0}, \mathbf{C}(\cdot, \cdot \,|\, \boldsymbol{\theta}))$. Any valid isotropic cross covariance function (see, e.g., Gelfand and Banerjee, 2010) can be used to construct $\mathbf{C}(\cdot, \cdot \,|\, \boldsymbol{\theta})$. To elucidate, let $\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ be the set of locations where the outcomes and predictors have been observed. This set may, but need not, intersect with the reference set $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ for the NNGP. Without loss of generality, we split up $\mathcal{T}$ into $\mathcal{S}^*$ and $\mathcal{U}$, where $\mathcal{S}^* = \mathcal{S} \cap \mathcal{T} = \{\mathbf{s}_{i_1}, \mathbf{s}_{i_2}, \ldots, \mathbf{s}_{i_r}\}$ with $\mathbf{s}_{i_j} = \mathbf{t}_j$ for $j = 1, 2, \ldots, r$ and $\mathcal{U} = \mathcal{T} \setminus \mathcal{S} = \{\mathbf{t}_{r+1}, \mathbf{t}_{r+2}, \ldots, \mathbf{t}_n\}$. Since $\mathcal{S} \cup \mathcal{T} = \mathcal{S} \cup \mathcal{U}$, we can completely specify the realizations of the NNGP in terms of the realizations of the parent process over $\mathcal{S}$ and $\mathcal{U}$, hierarchically, as $\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{w}_{\mathcal{S}} \sim N(\mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$ and $\mathbf{w}_{\mathcal{S}} \sim N(\mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{S}})$. For a full Bayesian specification, we further specify prior distributions on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and the τ_j^2's. For example, with customary prior specifications, we obtain the

joint distribution

$$p(\boldsymbol{\theta}) \times \prod_{j=1}^{q} IG(\tau_j^2 \,|\, a_{\tau_j}, b_{\tau_j}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \times N(\mathbf{w}_\mathcal{U} \,|\, \mathbf{B}_\mathcal{U}\mathbf{w}_\mathcal{S}, \mathbf{F}_\mathcal{U}) \\ \times N(\mathbf{w}_\mathcal{S} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_\mathcal{S}) \times \prod_{i=1}^{n} N(\mathbf{y}(\mathbf{t}_i) \,|\, \mathbf{X}(\mathbf{t}_i)'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t}_i)'\mathbf{w}(\mathbf{t}_i), \mathbf{D}) \;, \tag{2.10}$$

where $p(\boldsymbol{\theta})$ is the prior on $\boldsymbol{\theta}$ and $IG(\tau_j^2 \,|\, a_{\tau_j}, b_{\tau_j})$ denotes the Inverse-Gamma density.

2.3.2 Estimation and prediction

To describe a Gibbs sampler for estimating the parameters in (2.10), we define $\mathbf{y} = (\mathbf{y}(\mathbf{t}_1)', \mathbf{y}(\mathbf{t}_2)', \ldots, \mathbf{y}(\mathbf{t}_n)')'$, and $\mathbf{w}$ and $\boldsymbol{\epsilon}$ similarly. Also, we introduce $\mathbf{X} = [\mathbf{X}(\mathbf{t}_1) : \mathbf{X}(\mathbf{t}_2) : \ldots : \mathbf{X}(\mathbf{t}_n)]'$, $\mathbf{Z} = \text{diag}(\mathbf{Z}(\mathbf{t}_1)', \ldots, \mathbf{Z}(\mathbf{t}_n)')$, and $\mathbf{D}_n = \text{Cov}(\boldsymbol{\epsilon}) = \text{diag}(\mathbf{D}, \ldots, \mathbf{D})$. The full conditional distribution for $\boldsymbol{\beta}$ is $N(\mathbf{V}_\beta^*\boldsymbol{\mu}_\beta^*, \mathbf{V}_{\boldsymbol{\beta}}^*)$, where $\mathbf{V}_\beta^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}'\mathbf{D}_n^{-1}\mathbf{X})^{-1}$, $\boldsymbol{\mu}_\beta^* = (\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'\mathbf{D}_n^{-1}(\mathbf{y} - \mathbf{Z}\mathbf{w}))$. Inverse-Gamma priors for the τ_j^2's leads to conjugate full conditional distribution $IG(a_{\tau_j} + \frac{n}{2}, b_{\tau_j} + \frac{1}{2}(\mathbf{y}_{*j} - \mathbf{X}_{*j}\boldsymbol{\beta} - \mathbf{Z}_{*j}\mathbf{w})'(\mathbf{y}_{*j} - \mathbf{X}_{*j}\boldsymbol{\beta} - \mathbf{Z}_{*j}\mathbf{w})$ where $\mathbf{y}_{*j}$ refers to the $n \times 1$ vector containing the j^{th} co-ordinates of the $\mathbf{y}(\mathbf{t}_i)$'s, $\mathbf{X}_{*j}$ and $\mathbf{Z}_{*j}$ are the corresponding fixed and spatial effect covariate matrices respectively. For updating $\boldsymbol{\theta}$, we use a random walk Metropolis step with target density $p(\boldsymbol{\theta}) \times N(\mathbf{w}_\mathcal{S}|\mathbf{0}, \tilde{\mathbf{C}}_\mathcal{S}) \times N(\mathbf{w}_\mathcal{U} \,|\, \mathbf{B}_\mathcal{U}\mathbf{w}_\mathcal{S}, \mathbf{F}_\mathcal{U})$, where

$$\begin{aligned} N(\mathbf{w}_\mathcal{S} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_\mathcal{S}) &= \textstyle\prod_{i=1}^{k} N(\mathbf{w}(\mathbf{s}_i) \,|\, \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i}) \text{ and} \\ N(\mathbf{w}_\mathcal{U} \,|\, \mathbf{B}_\mathcal{U}\mathbf{w}_\mathcal{S}, \mathbf{F}_\mathcal{U}) &= \textstyle\prod_{i=r+1}^{n} N(\mathbf{w}(\mathbf{t}_i) \,|\, \mathbf{B}_{\mathbf{t}_i}\mathbf{w}_{N(\mathbf{t}_i)}, \mathbf{F}_{\mathbf{t}_i}) \end{aligned} \tag{2.11}$$

Each of the component densities under the product sign on the right hand side of (2.11) can be evaluated without any n-dimensional matrix operations rendering the NNGP suitable for efficient Metropolis (Hastings) block updates for $\boldsymbol{\theta}$.

The components of $\mathbf{w}_\mathcal{U} \,|\, \mathbf{w}_\mathcal{S}$ are independent. So we update $\mathbf{w}(\mathbf{t}_i) \,|\, \cdot \sim N(\mathbf{V}_{\mathbf{t}_i}\boldsymbol{\mu}_{\mathbf{t}_i}, \mathbf{V}_{\mathbf{t}_i})$ for $i = r+1, r+2, \ldots, n$ where $\boldsymbol{\mu}_{\mathbf{t}_i} = \mathbf{Z}(\mathbf{t}_i)\mathbf{D}^{-1}\left(\mathbf{y}(\mathbf{t}_i) - \mathbf{X}(\mathbf{t}_i)'\boldsymbol{\beta}\right) + \mathbf{F}_{\mathbf{t}_i}^{-1}\mathbf{B}_{\mathbf{t}_i}\mathbf{w}_{N(\mathbf{t}_i)}$ and $\mathbf{V}_{\mathbf{t}_i} = \left(\mathbf{Z}(\mathbf{t}_i)\mathbf{D}^{-1}\mathbf{Z}(\mathbf{t}_i)' + \mathbf{F}_{\mathbf{t}_i}^{-1}\right)^{-1}$. Finally, we update the components of $\mathbf{w}_\mathcal{S}$ individually. For any two locations $\mathbf{s}$ and $\mathbf{t}$ in $\mathcal{D}$, if $\mathbf{s} \in N(\mathbf{t})$ and is the l-th component of $N(\mathbf{t})$, i.e., say $\mathbf{s} = N(\mathbf{t})(l)$, then define $\mathbf{B}_{\mathbf{t},\mathbf{s}}$ as the $l \times l$ submatrix formed by columns $(l-1)q + 1, (l-1)q + 2, \ldots, lq$ of $\mathbf{B}_\mathbf{t}$. Let $U(\mathbf{s}_i) = \{\mathbf{t} \in \mathcal{S} \cup \mathcal{T} \,|\, \mathbf{s}_i \in N(\mathbf{t})\}$ and for every $\mathbf{t} \in U(\mathbf{s}_i)$ define, $\mathbf{a}_{\mathbf{t},\mathbf{s}_i} = \mathbf{w}(\mathbf{t}) - \sum_{\mathbf{s} \in N(\mathbf{t}), \mathbf{s} \neq \mathbf{s}_i} \mathbf{B}_{\mathbf{t},\mathbf{s}}\mathbf{w}(\mathbf{s})$. Then, for $i = 1, 2, \ldots, k$, we have the full

conditional $\mathbf{w}_{\mathbf{s}_i} \mid \cdot \sim N(\mathbf{V}_{\mathbf{s}_i}\boldsymbol{\mu}_{\mathbf{s}_i}, \mathbf{V}_{\mathbf{s}_i})$ where $\mathbf{V}_{\mathbf{s}_i} = (In(\mathbf{s}_i \in \mathcal{S}^*)\mathbf{Z}(\mathbf{s}_i)\mathbf{D}^{-1}\mathbf{Z}(\mathbf{s}_i)' + \mathbf{F}_{\mathbf{s}_i}^{-1} + \sum_{\mathbf{t}\in U(\mathbf{s}_i)} \mathbf{B}'_{\mathbf{t},\mathbf{s}_i}\mathbf{F}_{\mathbf{t}}^{-1}\mathbf{B}_{\mathbf{t},\mathbf{s}_i})^{-1}$, $\boldsymbol{\mu}_{\mathbf{s}_i} = In(\mathbf{s}_i \in \mathcal{S}^*)\mathbf{Z}(\mathbf{s}_i)\mathbf{D}^{-1}(\mathbf{y}(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta}) + \mathbf{F}_{\mathbf{s}_i}^{-1}\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)} + \sum_{\mathbf{t}\in U(\mathbf{s}_i)} \mathbf{B}'_{\mathbf{t},\mathbf{s}_i}\mathbf{F}_{\mathbf{t}}^{-1}\mathbf{a}_{\mathbf{t},\mathbf{s}_i}$ and $In(\cdot)$ denotes the indicator function. Hence, the $\mathbf{w}$'s can also be updated without requiring storage or factorization of any $n \times n$ matrices.

Turning to predictions, let $\mathbf{t}$ be a new location where we intend to predict $\mathbf{y}(\mathbf{t})$ given $\mathbf{X}(\mathbf{t})$ and $\mathbf{Z}(\mathbf{t})$. The Gibbs sampler for estimation also generates the posterior samples $\mathbf{w}_{\mathcal{S}} \mid \mathbf{y}$. So, if $\mathbf{t} \in \mathcal{S}$, then we simply get samples of $\mathbf{y}(\mathbf{t}) \mid \mathbf{y}$ from $N(\mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t}), \mathbf{D})$. If $\mathbf{t}$ is outside $\mathcal{S}$, then we generate samples of $\mathbf{w}(\mathbf{t}) \mid \cdot \sim N(\mathbf{V}_{\mathbf{t}}\boldsymbol{\mu}_{\mathbf{t}}, \mathbf{V}_{\mathbf{t}})$, where $\mathbf{V}_{\mathbf{t}} = \left(\mathbf{Z}(\mathbf{t})\mathbf{D}^{-1}\mathbf{Z}(\mathbf{t})' + \mathbf{F}_{\mathbf{t}}^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{t}} = \mathbf{Z}(\mathbf{t})\mathbf{D}^{-1}\left(\mathbf{y}(\mathbf{t}) - \mathbf{X}(\mathbf{t})'\boldsymbol{\beta}\right) + \mathbf{F}_{\mathbf{t}}^{-1}\mathbf{B}_{\mathbf{t}}\mathbf{w}_{N(\mathbf{t})}$, and subsequently generate posterior samples of $\mathbf{y}(\mathbf{t}) \mid \mathbf{y}$ similar to the earlier case.

2.3.3 Computational complexity

Implementing the NNGP model in Section 2.3.2 reveals that one entire pass of the Gibbs sampler can be completed without any large matrix operations. The only difference between (2.10) and a full geostatistical hierarchical model is that the spatial process is modeled as an NNGP prior as opposed to a standard GP. For comparisons, we offer rough estimates of the flop counts to generate $\boldsymbol{\theta}$ and $\mathbf{w}$ per iteration of the sampler. We express the computational complexity only in terms of the sample size n, size of the reference set k and the size of the neighbor sets m as other dimensions are assumed to be small. For all locations, $\mathbf{t} \in \mathcal{S} \cup \mathcal{T}$, $\mathbf{B}_{\mathbf{t}}$ and $\mathbf{F}_{\mathbf{t}}$ can be calculated using $O(m^3)$ flops. So, from (2.11) it is easy to see that $p(\boldsymbol{\theta} \mid \cdot)$ can be calculated using $O((n+k)m^3)$ flops. All subsequent calculations to generate a set of posterior samples for $\mathbf{w}$ and $\boldsymbol{\theta}$ require around $O((n+k)m^2)$ flops.

So, the total flop counts is of the order $(n+k)m^3$ and is ,therefore, linear in the total number of locations in $\mathcal{S} \cup \mathcal{T}$. This ensures scalability of the NNGP to large datasets. Compare this with a full GP model with a dense correlation matrix, which requires $O(n^3)$ flops for updating $\mathbf{w}$ in each iteration. Simulation results in Section 2.5.1 indicate that NNGP models with usually very small values of m (≈ 10) provides inference almost indistinguishable to full geostatistical models. Therefore, for large n, this linear flop count is drastically less. Also, linearity with respect to k ensures a feasible implementation even for $k \approx n$. This offers substantial improvement over low rank models where the computational cost is quadratic in the number of "knots," limiting

the size of the set of knots. Also, both the full geostatistical and the predictive process models require storage of the $n \times n$ distance matrix, which can potentially exhaust storage resources for large datasets. An NNGP model only requires the distance matrix between neighbors for every location, thereby storing $n+k$ small matrices, each of order $m \times m$. Hence, NNGP accrues substantial computational benefits over existing methods for very large spatial datasets and may be the only feasible option for fully model-based inference in certain cases, as seen in the forestry data example (Section 2.5.3).

2.3.4 Model comparison and choice of $\mathcal{S}$ and m

As elaborated in Section 2.2, given any parent Gaussian process and *any* fixed reference set of locations $\mathcal{S}$, we can construct a valid NNGP. The resulting finite dimensional likelihoods of the NNGP depend upon the choice of the reference set $\mathcal{S}$ and the size of each $N(\mathbf{s}_i)$, i.e., m. Choosing the reference set is similar to selecting the knots for a predictive process. Unlike the number of "knots" in low rank models, the number of points in $\mathcal{S}$ do not thwart computational scalability. From Section 2.3.3, we observe that the flop count in an NNGP model only increases linearly with the size of $\mathcal{S}$. Hence, the number of locations in $\mathcal{S}$ can, in theory, be large and this provides a lot of flexibility in choosing $\mathcal{S}$.

Points over a grid across the entire domain seem to be a plausible choice for $\mathcal{S}$. For example, we can construct a large $\mathcal{S}$ using a dense grid to improve performance without adversely affecting computational costs. Another, perhaps even simpler, option for large datasets is to simply fix $\mathcal{S} = \mathcal{T}$, the set of observed locations. Since the NNGP is a legitimate process for any fixed $\mathcal{S}$, this choice is legitimate and it reduces computational costs even further by avoiding additional sampling of $\mathbf{w}_{\mathcal{U}}$ in the Gibbs sampler. Our empirical investigations (see Section 2.5.1) reveal that choosing $\mathcal{S} = \mathcal{T}$ deliver inference almost indistinguishable from choosing $\mathcal{S}$ to be a grid over the domain for large datasets.

Stein et al. (2004) and Eidsvik et al. (2014) proposed using a sandwich variance estimator for evaluating the inferential abilities of neighbor-based pseudo-likelihoods. Shaby (2012) developed a post sampling sandwich variance adjustment for posterior credible intervals of the parameters for quasi-Bayesian approaches using pseudo-likelihoods. However, all these adjustments concede accrual of additional computational costs. Also, the asymptotic results used to obtain the sandwich variance estimators are based on

assumptions which are hard to verify in spatial settings with irregularly placed data points. Moreover, we view the NNGP as an independent model for fitting the data and not as an approximation to the original GP. Hence, we refrain from such sandwich variance adjustments. Instead, we can simply use any standard model comparison metrics such as DIC (Spiegelhalter et al., 2002), GPD (Gelfand and Ghosh, 1998) or RMSPE(RMSECV) (Yeniay and Goktas, 2002) to compare the performance of NNGP and any other candidate model. The same model comparison metrics are also used for selecting m. However, as we illustrate later in Section 2.5.1, usually a small value of m between 10 to 15 produces performance at par with the full geostatistical model. While larger m may be beneficial for massive datasets, perhaps under a different design scheme, it is still going to be much smaller than the number of knots required in low rank models (see Section 2.5.1).

2.4 Alternate NNGP models and algorithms

2.4.1 Block update of $\mathbf{w}_{\mathcal{S}}$ using sparse Cholesky

The Gibbs' sampling algorithm detailed in Section 2.3.2 is extremely efficient for large datasets with linear flop counts per iteration. However, it can sometimes experience slow convergence issues due to sequential updating of the elements in $\mathbf{w}_{\mathcal{S}}$. An alternative to sequential updating is to perform block updates of $\mathbf{w}_{\mathcal{S}}$. We choose $\mathcal{S} = \mathcal{T}$ so that $\mathbf{s}_i = \mathbf{t}_i$ for all $i = 1, 2, \ldots, n$ and we denote $\mathbf{w}_{\mathcal{S}} = \mathbf{w}_{\mathcal{T}}$ by $\mathbf{w}$. Then,

$$\mathbf{w}|\cdot \sim N(\mathbf{V}_{\mathcal{S}}\mathbf{Z}'\mathbf{D}_n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{V}_{\mathcal{S}}) \ , \quad \text{where} \quad \mathbf{V}_{\mathcal{S}} = (\mathbf{Z}'\mathbf{D}_n^{-1}\mathbf{Z} + \tilde{\mathbf{C}}_{\mathcal{S}}^{-1})^{-1} \ . \tag{2.12}$$

Recall that $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ is sparse. Since $\mathbf{Z}$ and $\mathbf{D}_n$ are block diagonal, $\mathbf{V}_{\mathcal{S}}^{-1}$ retains the sparsity of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$. So, a sparse Cholesky factorization of $\mathbf{V}_{\mathcal{S}}^{-1}$ will efficiently produce the Cholesky factors of $\mathbf{V}_{\mathcal{S}}$. This will facilitate block updating of $\mathbf{w}$ in the Gibbs sampler.

2.4.2 NNGP models for the response

Another possible approach involves NNGP models for the response $\mathbf{y}(\mathbf{s})$. If $\mathbf{w}(\mathbf{s})$ is a Gaussian Process, then so is $\mathbf{y}(\mathbf{s}) = \mathbf{Z}(\mathbf{s})'\mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}$ (without loss of generality we assume $\beta = \mathbf{0}$). One can directly use the NNGP specification for $\mathbf{y}(\mathbf{s})$ instead of $\mathbf{w}(\mathbf{s})$. That is, we derive $\mathbf{y}(\mathbf{s}) \sim NNGP(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}(\cdot, \cdot))$ from the parent Gaussian process

$GP(\mathbf{0}, \boldsymbol{\Sigma}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$. The Gibbs sampler analogous to Section 2.3 now enjoys the additional advantage of avoiding full conditionals for $\mathbf{w}$. This results in a Bayesian analogue for Vecchia (1988) and Stein et al. (2004) but precludes inference on the spatial residual surface $\mathbf{w}(\mathbf{s})$. Modeling $\mathbf{w}(\mathbf{s})$ provides additional insight into residual spatial contours and is often important in identifying lurking covariates or eliciting unexplained spatial patterns. Vecchia (1992) used the nearest neighbor approximation on a spatial model for observations ($\mathbf{y}$) with independent measurement error (nuggets) in addition to the usual spatial component ($\mathbf{w}$). However, it may not be possible to recover $\mathbf{w}$ using this approach. For example, a univariate stationary process $\mathbf{y}(\mathbf{s})$ with a nugget effect can be decomposed as $\mathbf{y}(\mathbf{s}) = \mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$ (letting $\boldsymbol{\beta} = \mathbf{0}$) for some $\mathbf{w}(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$ and white noise process $\boldsymbol{\epsilon}(\mathbf{s})$. If $\mathbf{y} = \mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{w} \sim N(\mathbf{0}, \mathbf{C})$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$, then $\text{Cov}(\mathbf{y}) = \mathbf{C} + \tau^2\mathbf{I} = \boldsymbol{\Sigma}$, all eigenvalues of $\boldsymbol{\Sigma}$ are greater than τ^2 and $\text{Cov}(\mathbf{w} \,|\, \mathbf{y}) = \tau^2\mathbf{I}_n - \tau^4\boldsymbol{\Sigma}^{-1}$. For $\mathbf{y}(\mathbf{s}) \sim NNGP(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}(\cdot,\cdot))$, however, the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ may be less than τ^2, so $\tau^2\mathbf{I}_n - \tau^4\tilde{\boldsymbol{\Sigma}}^{-1}$ need not be positive definite for every $\tau^2 > 0$ and $p(\mathbf{w} \,|\, \mathbf{y})$ is no longer well-defined.

A different model is obtained by using an NNGP prior for $\mathbf{w}$, as in (2.10), and then integrating out $\mathbf{w}$. The resulting likelihood is $N(\mathbf{y} \,|\, \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{\Sigma}_y = \mathbf{Z}\tilde{\mathbf{C}}_{\mathcal{S}}\mathbf{Z}' + \mathbf{D}_n$ and the Bayesian specification is completed using priors on $\boldsymbol{\beta}$, τ_j^2's and $\boldsymbol{\theta}$ as in (2.10). This model drastically reduces the number of variables in the Gibbs sampler, while preserving the nugget effect in the parent model. We can generate the full conditionals for the parameters in the marginalized model as follows: $\boldsymbol{\beta} |\ \mathbf{y}, \boldsymbol{\phi} \sim N((\mathbf{V}_\beta^{-1} + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-1}(\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y}), (\mathbf{V}_\beta^{-1} + \mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-1})$. It is difficult to factor out τ_j^2's from $\boldsymbol{\Sigma}_y^{-1}$, so conjugacy is lost with respect to any standard prior. Metropolis block updates for $\boldsymbol{\theta}$ are feasible for any tractable prior $p(\boldsymbol{\theta})$. This involves computing $\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{X}$, $\mathbf{X}'\boldsymbol{\Sigma}_y^{-1}\mathbf{y}$ and $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Since $\boldsymbol{\Sigma}_y^{-1} = \mathbf{D}_n^{-1} - \mathbf{D}_n^{-1}\mathbf{Z}(\tilde{\mathbf{C}}_{\mathcal{S}}^{-1} + \mathbf{Z}'\mathbf{D}_n^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_n^{-1} = \mathbf{D}_n^{-1} - \mathbf{D}_n^{-1}\mathbf{Z}\mathbf{V}_{\mathcal{S}}\mathbf{Z}'\mathbf{D}_n^{-1}$, where $\mathbf{V}_{\mathcal{S}}$ is given by (2.12), a sparse Cholesky factorization of $\mathbf{V}_{\mathcal{S}}^{-1}$ will be beneficial. We draw posterior samples for $\mathbf{w}$ from $p(\mathbf{w} \,|\, \mathbf{y}) = \int p(\mathbf{w} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \{\tau_j^2\}, \mathbf{y}) p(\boldsymbol{\theta}, \boldsymbol{\beta}, \{\tau_j^2\} \,|\, \mathbf{y})$ using composition sampling—we draw $\mathbf{w}^{(g)}$ from $p(\mathbf{w} \,|\, \boldsymbol{\theta}^{(g)}, \boldsymbol{\beta}^{(g)}, \{\tau_j^{2(g)}\}, \mathbf{y})$ one-for-one for each sampled parameter.

Using block updates for $\mathbf{w}_{\mathcal{S}}$ in (2.10) and fitting the marginalized version of (2.10) both require an efficient sparse Cholesky solver for $\mathbf{V}_{\mathcal{S}}^{-1}$. Note that computational expenses for most sparse Cholesky algorithms depend on the precise nature of the sparse

structure (mostly on the bandwidth) of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ (see, e.g. Davis, 2006). The number of flops required for Gibbs sampling and prediction in this marginalized model depends upon the sparse structure of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ and may, sometimes, heavily exceed the linear usage achieved by the unmarginalized model with individual updates for $\mathbf{w}_i$. Therefore, a prudent choice of the precise fitting algorithms should be based on the sparsity structure of $\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$ for the given dataset.

2.4.3 Spatiotemporal and GLM versions

In spatiotemporal settings where we seek spatial interpolation at discrete time-points (e.g., weekly, monthly or yearly data), we write the response (possibly vector-valued) as $\mathbf{y}_t(\mathbf{s})$ and the random effects as $\mathbf{w}_t(\mathbf{s})$. One could, for example, envision that the data arise as a time series of spatial processes, i.e., there is a time series at each location. An alternative scenario is cross-sectional data being collected at a set of locations associated with each time point and these locations can differ from time point to time point. Desired inference includes spatial interpolation for each time point. Spatial dynamic models incorporating the NNGP are easily formulated as below:

$$\begin{aligned} \mathbf{y}_t(\mathbf{s}) &= \mathbf{X}_t(\mathbf{s})'\boldsymbol{\beta}_t + \mathbf{u}_t(s) + \boldsymbol{\epsilon}_t(\mathbf{s}),\ \boldsymbol{\epsilon}_t(\mathbf{s}) \overset{iid}{\sim} N(0, D) \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}_\eta),\ \boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, \boldsymbol{\Sigma}_0) \\ \mathbf{u}_t(\mathbf{s}) &= \mathbf{u}_{t-1}(\mathbf{s}) + \mathbf{w}_t(\mathbf{s}),\ \mathbf{w}_t(\mathbf{s}) \overset{ind}{\sim} NNGP(\mathbf{0}, \tilde{\mathbf{C}}(\cdot,\cdot \,|\, \boldsymbol{\theta}_t)) \ . \end{aligned} \tag{2.13}$$

Thus, one retains exactly the same structure of process-based spatial dynamic models, e.g., as in Gelfand et al. (2005a), and simply replaces the independent Gaussian process priors for $\mathbf{w}_t(\mathbf{s})$ with independent NNGP's to achieve computational tractability.

The above is illustrative of how attractive and extremely convenient the NNGP is for model building. One simply writes down the parent model and subsequently replaces the full GP with an NNGP. Being a well-defined process, the NNGP ensures a valid spatial dynamic model. Similarly NNGP versions of dynamic spatiotemporal Kalman-filtering (Wikle and Cressie, 1999, as, e.g., in) can be constructed.

Handling non-Gaussian (e.g., binary or count) data is also straightforward using spatial generalized linear models (GLM's) (Diggle et al., 1998; Lin et al., 2000; Kammann and Wand, 2003; Banerjee et al., 2014). Here, the NNGP provides structured dependence for random effects at the second stage. First, we replace $\mathrm{E}[\mathbf{y}(\mathbf{t})]$ in (2.9)

with $g(E(\mathbf{y}(\mathbf{t})))$ where $g(\cdot)$ is a suitable link function such that $\boldsymbol{\eta}(\mathbf{t}) = g(E(\mathbf{y}(\mathbf{t}))) = \mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t})'\mathbf{w}(\mathbf{t})$. In the second stage, we model the $\mathbf{w}(\mathbf{t})$ as an NNGP. The benefits of the algorithms in Sections 2.3.2 and 2.3.3 still hold, but some of the alternative algorithms in Section 2.4 may not apply. For example, we do obtain tractable marginalized likelihoods by integrating out the spatial effects.

2.5 Illustrations

We conduct simulation experiments and analyze a large forestry dataset. Posterior inference for subsequent analysis were based upon three chains of 25000 iterations (with a burn-in of 5000 iterations). All the samplers were programmed in `C++` and leveraged Intels Math Kernel Library's (MKL) threaded `BLAS` and `LAPACK` routines for matrix computations on a Linux workstation with 384 GB of RAM and two Intel Nehalem quad-Xeon processors.

2.5.1 Simulation experiment

We generated observations using 2500 locations within a unit square domain from the model (2.9) with $q = l = 1$ (univariate outcome), $p = 2$, $\mathbf{Z}(\mathbf{t})' = 1$ (scalar), the spatial covariance matrix $\mathbf{C}(\boldsymbol{\theta}) = \sigma^2\mathbf{R}(\boldsymbol{\phi})$, where $\mathbf{R}(\boldsymbol{\phi})$ is a $n \times n$ correlation matrix, and $\mathbf{D} = \tau^2$ (scalar). The model included an intercept and a covariate $\mathbf{x}_1$ drawn from $N(0, 1)$. The (i, j)th element of $\mathbf{R}(\boldsymbol{\phi})$ was calculated using the Matérn function

$$\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\phi}) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(||\mathbf{t}_i - \mathbf{t}_j||\phi)^{\nu}\mathcal{K}_{\nu}(||\mathbf{t}_i - \mathbf{t}_j||\phi);\ \phi > 0, \nu > 0, \tag{2.14}$$

where $||\mathbf{t}_i - \mathbf{t}_j||$ is the Euclidean distance between locations $\mathbf{t}_i$ and $\mathbf{t}_j$, $\boldsymbol{\phi} = (\phi, \nu)$ with ϕ controlling the decay in spatial correlation and ν controlling the process smoothness, Γ is the usual Gamma function while $\mathcal{K}_{\nu}$ is a modified Bessel function of the second kind with order ν (Stein, 1999) Evaluating the Gamma function for each matrix element within each iteration requires substantial computing time and can obscure differences in sampler run times; hence, we fixed ν at 0.5 which reduces (2.14) to the exponential correlation function. The first column in Table 2.1 gives the *true* values used to generate the responses. Figure 2.2(a) illustrates the $w(\mathbf{t})$ surface interpolated over the domain.

We then estimated the following models from the full data: *i*) the full Gaussian Process (*Full GP*); *ii*) the NNGP with $m = \{1, 2, \ldots, 25\}$ for $\mathcal{S} \neq \mathcal{T}$ and $\mathcal{S} = \mathcal{T}$, and; *iii*) a Gaussian Predictive Process (GPP) model (Banerjee et al., 2008) with 64 knots placed on a grid over the domain. For the NNGP with $\mathcal{S} \neq \mathcal{T}$ we considered 2000 randomly placed reference locations within the domain. The 64 knot GPP was chosen because its computing time was comparable to that of NNGP models. We used an efficient marginalized sampling algorithm for the Full GP and GPP models as implemented in the `spBayes` package in `R` (Finley et al., 2013). All the models were trained using 2000 of the 2500 observed locations, while the remaining 500 observations were withheld to assess predictive performance.

For all models, the intercept and slope regression parameters, β_0 and β_1, were given *flat* prior distributions. The variance components σ^2 and τ^2 were assigned inverse-Gamma $IG(2, 1)$ and $IG(2, 0.1)$ priors, respectively, and the spatial decay ϕ received a uniform prior $U(3, 30)$, which corresponds to a spatial range between approximately 0.1 and 1 units.

Parameter estimates and performance metrics for the NNGP (with $m = 10$ and $m = 20$), GPP, and the Full GP models are provided in Table 2.1. All model specifications produce similar posterior median and 95% credible intervals estimates, with the exception of ϕ in the 64 knot GPP model. Larger values of DIC and D suggest that the GPP model does not fit the data as well as the NNGP and Full GP models. The NNGP $\mathcal{S} = \mathcal{T}$ models provide DIC, GPD scores that are comparable to those of the Full GP model. These fit metrics suggest the NNGP $\mathcal{S} \neq \mathcal{T}$ models provide better fit to the data than that achieved by the full GP model which is probably due to overfitting caused by a very large reference set $\mathcal{S}$. The last row in Table 2.1 shows computing times in minutes for one chain of 25000 iterations reflecting on the enormous computational gains of NNGP models over full GP model.

Turning to out-of-sample predictions, the Full model's RMSPE and mean width between the upper and lower 95% posterior predictive credible interval is 1.2 and 2.12, respectively. As seen in Figure 2.1, comparable RMSPE and mean interval width for the NNGP $\mathcal{S} = \mathcal{T}$ model is achieved within $m \approx 10$. There are negligible difference between the predictive performances of the NNGP $\mathcal{S} \neq \mathcal{T}$ and $\mathcal{S} = \mathcal{T}$ models. Both the NNGP and Full GP model have better predictive performance than the Predictive Process

Table 2.1: Univariate synthetic data analysis parameter estimates and computing time in minutes for NNGP and full GP models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

		NNGP ($\mathcal{S} \neq \mathcal{T}$)		NNGP ($\mathcal{S} = \mathcal{T}$)	
	True	$m = 10, k = 2000$	$m = 20, k = 2000$	$m = 10$	$m = 20$
β_0	1	0.99 (0.71, 1.48)	1.02 (0.73, 1.49)	1.00 (0.62, 1.31)	1.03 (0.65, 1.34)
β_1	5	5.00 (4.98, 5.03)	5.01 (4.98, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)
σ^2	1	1.09 (0.89, 1.49)	1.04 (0.85, 1.40)	0.96 (0.78, 1.23)	0.94 (0.77, 1.20)
τ^2	0.1	0.07 (0.04, 0.10)	0.07 (0.04, 0.10)	0.10 (0.08, 0.13)	0.10 (0.08, 0.13)
ϕ	12	11.81 (8.18, 15.02)	12.21 (8.83, 15.62)	12.93 (9.70, 16.77)	13.36 (9.99, 17.15)
p_D	–	1491.08	1478.61	1243.32	1249.57
DIC	–	1856.85	1901.57	2390.65	2377.51
G	–	33.67	35.68	77.84	76.40
P	–	253.03	259.13	340.40	337.88
D	–	286.70	294.82	418.24	414.28
RMSPE	–	1.22	1.22	1.2	1.2
95% CI cover %	–	97.2	97.2	97.6	97.6
95% CI width	–	2.19	2.18	2.13	2.12
Time	–	14.2	47.08	9.98	33.5

		Predictive Process	Full
	True	64 knots	Gaussian Process
β_0	1	1.30 (0.54, 2.03)	1.03 (0.69, 1.34)
β_1	5	5.03 (4.99, 5.06)	5.01 (4.99, 5.03)
σ^2	1	1.29 (0.96, 2.00)	0.94 (0.76, 1.23)
τ^2	0.1	0.08 (0.04, 0.13)	0.10 (0.08, 0.12)
ϕ	12	**5.61 (3.48, 8.09)**	13.52 (9.92, 17.50)
p_D	–	1258.27	1260.68
DIC	–	13677.97	2364.80
G	–	1075.63	74.80
P	–	200.39	333.27
D	–	1276.03	408.08
RMSPE	–	1.68	1.2
95% CI cover %	–	95.6	97.6
95% CI width	–	2.97	2.12
Time	–	43.36	560.31

models when the number of knots is small, e.g., 64. All models showed appropriate 95% credible interval coverage rates.

Figures 2.2(b-f) illustrate the posterior median estimates of the spatial random effects from the Full GP, NNGP ($\mathcal{S} = \mathcal{T}$) with $m = 10$ and $m = 20$, NNGP ($\mathcal{S} \neq \mathcal{T}$) with $m = 10$ and GPP models. These surfaces can be compared to the *true* surface depicted in Figure 2.2(a). This comparison shows: *i*) the NNGP models closely approximates the true surface and that estimated by the Full GP model, and; *ii*) the reduced rank predictive process model based on 64 knots greatly smooths over small-scale patterns. This last observation highlights one of the major criticisms of reduced rank models Stein (2014) and illustrates why these models often provide compromised predictive performance when the true surface has fine spatial resolution details.

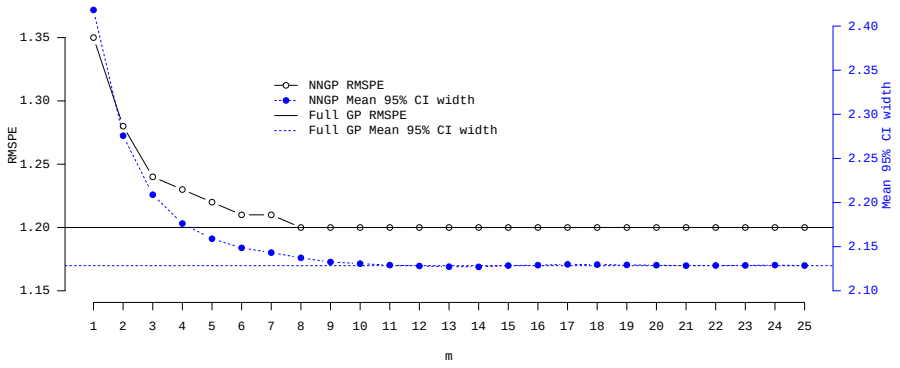
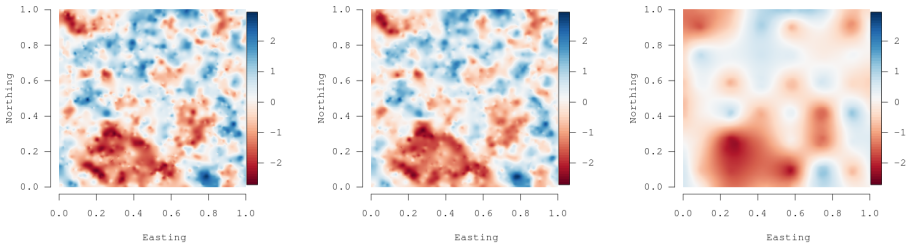
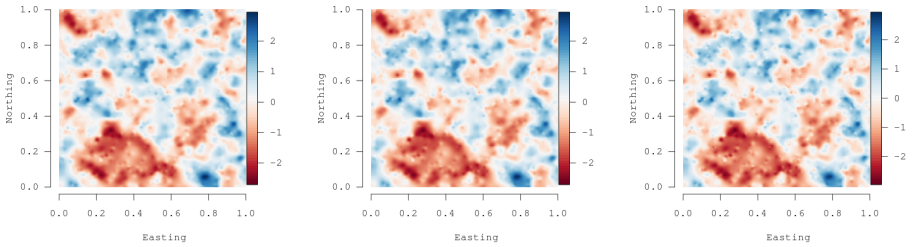


Figure 2.1: Choice of m in NNGP models: Out-of-sample Root Mean Squared Prediction Error (RMSPE) and mean width between the upper and lower 95% posterior predictive credible intervals for a range of m for the univariate synthetic data analysis



(a) True $\mathbf{w}$ (b) Full GP (c) GPP 64 knots



(d) NNGP $(\mathcal{S} = \mathcal{T})$ $m = 10$ (e) NNGP $(\mathcal{S} = \mathcal{T})$ $m = 20$ (f) NNGP $(\mathcal{S} \neq \mathcal{T})$ $m = 10$

Figure 2.2: Univariate synthetic data analysis: Interpolated surfaces of the true spatial random effects and posterior median estimates for different models

Overall, we see the clear computational advantage of the NNGP over the Full GP model, and both inferential and computational advantage over the GPP model.

2.5.2 Robustness of NNGP to ordering of locations

We conduct a simulation experiment demonstrating the robustness of NNGP to the ordering of the locations. We generate the data for $n = 2500$ locations using the model in Section 2.5.1. However instead of a square domain we choose a long skinny domain (see Figure 2.3(a)) which can bring out possible sensitivity to ordering due to scale disparity between the x and y axes. We use three different orderings for the locations: ordering by x-coordinates, by y-coordinates and by the function $f(x, y) = x + y$.

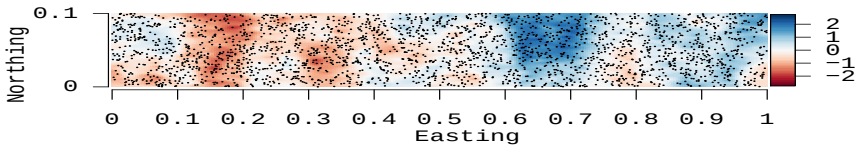
Table 2.2 demonstrates that the point estimates and the 95% credible intervals for

Table 2.2: Univariate synthetic data analysis parameter estimates and computing time in minutes for NNGP m=10 and full GP models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

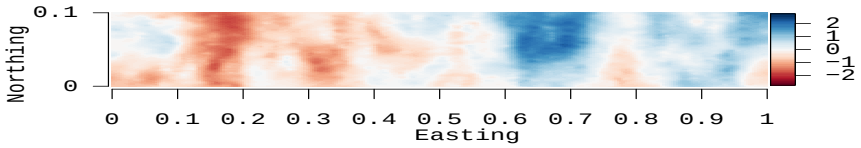
			NNGP ($\mathcal{S} = \mathcal{T}$)		
	True	Full Gaussian Process	Order by y-coordinates	Order by x-coordinates	Order by $x + y$-coordinates
σ^2	1	0.64 (0.41, 1.30)	0.71 (0.45, 1.53)	0.76 (0.48, 1.50)	0.72 (0.46, 1.44)
τ^2	0.1	0.11 (0.10, 0.12)	0.11 (0.10, 0.11)	0.11 (0.10, 0.12)	0.11 (0.10, 0.12)
ϕ	6	8.26 (4.06, 13.41)	8.29 (3.56, 12.88)	7.13 (3.41, 11.27)	7.50 (3.60, 11.91)

the process parameters from all three NNGP models are extremely consistent with the estimates from the full Gaussian process model.

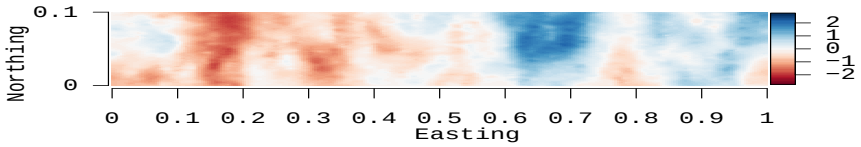
Posterior estimates of the spatial residual surface from the different models are shown in Figure 2.3. Again, the impact of the different ordering is negligible. In Figure 2.4, we plotted the difference between the posterior estimates of the random effects of the true GP and NNGP for all 3 orderings. It was seen that this difference was negligible compared to the difference between the true spatial random effects and full GP estimates. This shows the inference obtained from the NNGP (using any ordering) closely emulates the corresponding full GP inference.



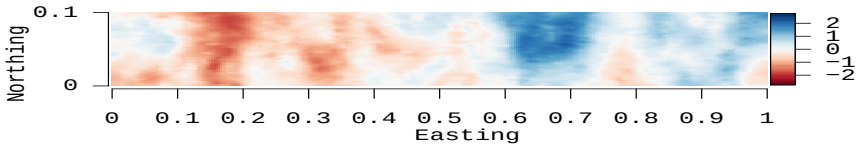
(a) True $\mathbf{w}$



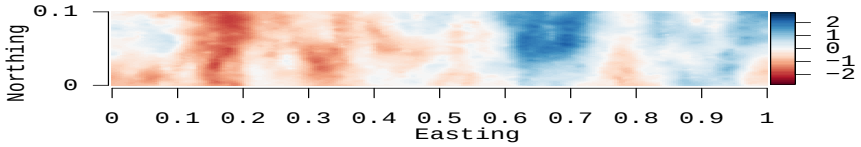
(b) Full GP



(c) NNGP order by y-coordiantes

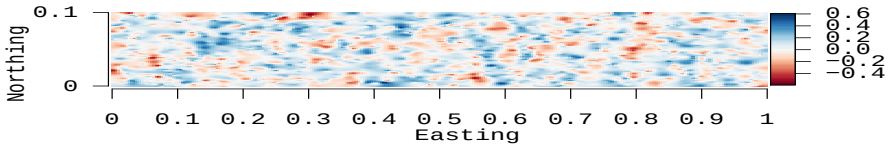


(d) NNGP order by x-coordiantes

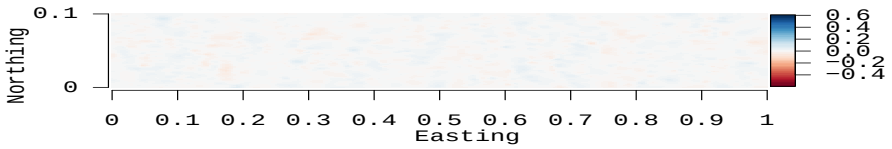


(e) NNGP order by $x + y$-coordiantes

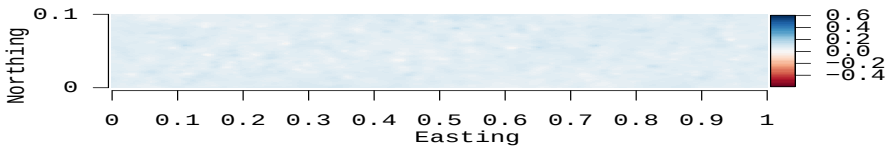
Figure 2.3: Robustness of NNGP to ordering: Figures (a) and (b) show interpolated surfaces of the true spatial random effects and posterior median estimates for full geostatistical model respectively. Figures (c), (d), and (e) show interpolated surfaces of the posterior median estimates for NNGP model with $\mathcal{S} = \mathcal{T}$, $m = 10$, and alternative coordinate ordering. Corresponding true and estimated process parameters are given in Table 2.2.



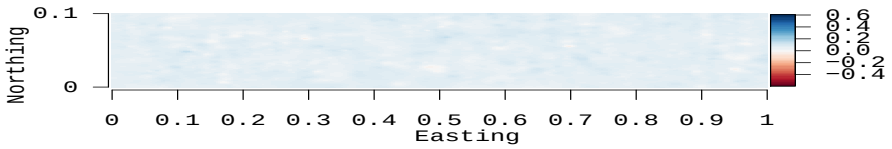
(a) True $\mathbf{w}$− Full GP $\widehat{\mathbf{w}}$



(b) Full GP $\widehat{\mathbf{w}}$− NNGP (order by x) $\widehat{\mathbf{w}}$



(c) Full GP $\widehat{\mathbf{w}}$− NNGP (order by y) $\widehat{\mathbf{w}}$



(d) Full GP $\widehat{\mathbf{w}}$− NNGP (order by $x + y$) $\widehat{\mathbf{w}}$

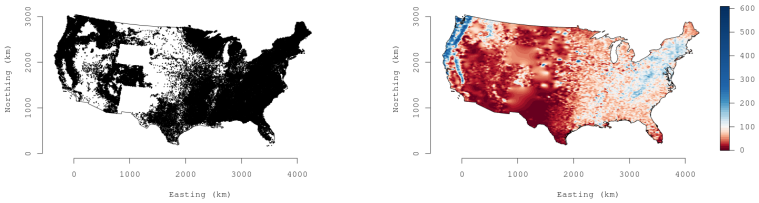
Figure 2.4: Difference between Full GP and NNGP estimates of spatial effects: Figure (a) shows the difference between the true spatial random effects and the full GP posterior median estimates. Figures (b), (c) and (d) plots the difference between posterior median estimates of full GP and NNGP ordered by x, y and $x + y$ co-ordinates respectively. All the figures are in the same color scale.

2.5.3 Forest biomass data analysis

Information about the spatial distribution of forest biomass is needed to support global, regional, and local scale decisions, including assessment of current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation. In the United States, the Forest Inventory and Analysis (FIA) program of the USDA Forest Service collects the data needed to support these assessments. The program has established field plot centers in permanent locations using a sampling design that produces an equal probability sample (Bechtold and Patterson, 2005). Field crews recorded stem measurements for all trees with diameter at breast height (DBH); 1.37 m above the forest floor) of 12.7 cm or greater. Given these data, established allometric equations were used to estimate each plot's forest biomass. For the subsequent analysis, plot biomass was scaled to metric tons per ha then square root transformed. The transformation ensures that back transformation of subsequent predicted values have support greater than zero and helps to meet basic regression models assumptions.

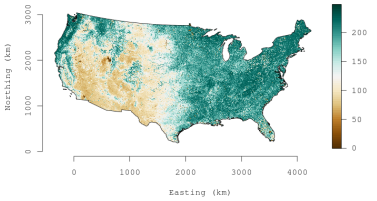
Figure 2.5(a) illustrates the georeferenced forest inventory data consisting of $114,371$ forested FIA plots measured between 1999 and 2006 across the conterminous United States. The two blocks of missing observations in the Western and Southwestern United States correspond to Wyoming and New Mexico, which have not yet released FIA data. Figure 2.5(b) shows a deterministic interpolation of forest biomass observed on the FIA plots. Dark blue indicates high forest biomass, which is primarily seen in the Pacific Northwest, Western Coastal ranges, Eastern Appalachian Mountains, and in portions of New England. In contrast, dark red indicates regions where climate or land use limit vegetation growth.

A July 2006 Normalized Difference Vegetation Index (NDVI) image from the MODerate resolution Imaging Spectroradiometer (MODIS); `http://glcf.umd.edu/data/ndvi`) sensor was used as a single predictor. NDVI is calculated from the visible and near-infrared light reflected by vegetation, and can be viewed as a measure of greenness. In this image, Figure 2.5(c), dark green corresponds to dense vegetation whereas brown identifies regions of sparse or no vegetation, e.g., in the Southwest. NDVI is commonly used as a covariate in forest biomass regression models, see, for e.g., Zhang and Kondraguanta (2006). Results from these and similar studies show a positive linear relationship between forest biomass and NDVI. The strength of this relationship, however,

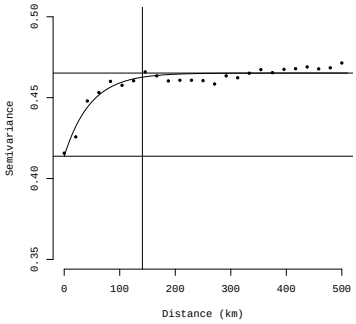


(a) Observed locations

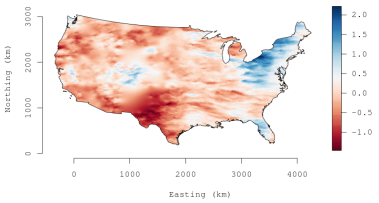
(b) Observed biomass



(c) NDVI



(d) Non-spatial model residuals



(e) SVI $\beta_0(\mathbf{t})$

Figure 2.5: Forest biomass data analysis: (a) locations of observed biomass, (b) interpolated biomass response variable, (c) NDVI regression covariate, (d) variogram of non-spatial model residuals, and (e) surface of the SVI model random spatial effects posterior medians. Following our FIA data sharing agreement, plot locations depicted in (a) have been "fuzzed" to hide the true coordinates.

varies by forest tree species composition, age, canopy structure, and level of reflectance. We expect a space-varying relationship between biomass and NDVI, given tree species composition and disturbance regimes generally exhibit strong spatial dependence across forested landscapes.

The ∼38 gigabytes of memory in our workstation was insufficient for storage of distance matrices required to fit a Full GP or GPP model. Subsequently, we explore the relationship between forest biomass and NDVI using a non-spatial model, a NNGP space-varying intercept (SVI) model (i.e., $q = l = 1$ and $\mathbf{Z}(\mathbf{t}) = 1$) in (2.9), and a NNGP spatially-varying coefficients (SVC) regression model with $l = 1$, $q = p = 2$ and $\mathbf{Z}(\mathbf{t}) = \mathbf{X}(\mathbf{t})$ in (2.9). The reference sets for the NNGP models were again the observed locations and m was chosen to be 5 or 10. The parent process $\mathbf{w}(\mathbf{t})$ is a bivariate Gaussian process with a isotropic cross-covariance specification $\mathbf{C}(\mathbf{t}_i, \mathbf{t}_j \,|\, \boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\Gamma}(\boldsymbol{\phi})\mathbf{A}'$, where $\mathbf{A}$ is 2×2 lower-triangular with positive diagonal elements, $\boldsymbol{\Gamma}$ is 2×2 diagonal with $\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\phi}_b)$ (defined in (2.14)) as the b^{th} diagonal entry, $b = 1, 2$ and $\boldsymbol{\phi}_b = (\phi_b, \nu_b)'$ (see, e.g., Gelfand and Banerjee, 2010).

For all models, the intercept and slope regression parameters were given *flat* prior distributions. The variance components τ^2 and σ^2 were assigned inverse-Gamma $IG(2, 1)$ priors, the SVC model cross-covariance matrix $\mathbf{A}\mathbf{A}'$ was given an inverse-Wishart $IW(3, 0.1)$, and the Matérn spatial decay and smoothness parameters received uniform prior supports $U(0.01, 3)$ and $U(0.1, 2)$, respectively. These prior distributions on ϕ and ν correspond to support between approximately 0.5 and 537 km. Candidate models are assessed using the metrics described in Section 2.3.4, inference drawn from mapped estimates of the regression coefficients, and out-of-sample prediction.

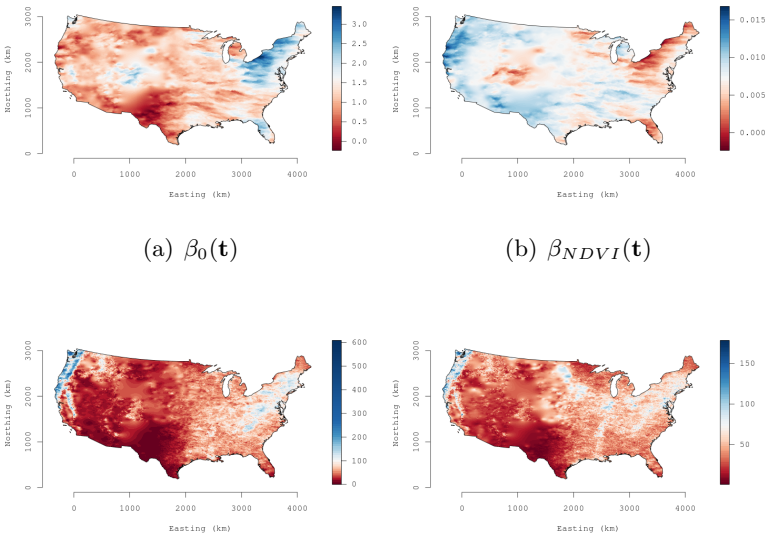
Parameter estimates and performance metrics for NNGP with $m = 5$ are shown in Table 2.3. The corresponding numbers for $m = 10$ were similar. Relative to the spatial models, the non-spatial model has higher values of DIC and D which suggests NDVI alone does not adequately capture the spatial structure of forest biomass. This observation is corroborated using a variogram fit to the non-spatial model's residuals, Figure 2.5(d). The variogram shows a nugget of ∼0.42, partial sill of ∼0.05, and range of ∼150 km. This residual spatial dependence is apparent when we map the SVI model spatial random effects as shown in Figure 2.5(e). This map, and the estimate of a non-negligible spatial variance σ^2 in Table 2.3, suggests the addition of a spatial random

effect was warranted and helps satisfy the model assumption of uncorrelated residuals.

Table 2.3: Forest biomass data analysis parameter estimates and computing time in hours for candidate models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

		NNGP	NNGP
	Non-spatial	Space-varying intercept	Space-varying coefficients
β_0	1.043 (1.02, 1.065)	1.44 (1.39, 1.48)	1.23 (1.20, 1.26)
β_{NDVI}	0.0093 (0.009, 0.0095)	0.0061 (0.0059, 0.0062)	0.0072 (0.0071, 0.0074)
σ^2	–	0.16 (0.15, 0.17)	–
$\mathbf{AA}'_{1,1}$	–	–	0.24 (0.23, 0.24)
$\mathbf{AA}'_{2,1}$	–	–	-0.00088 (-0.00093, -0.00083)
$\mathbf{AA}'_{2,2}$	–	–	0.0000052 (0.0000047, 0.0000056)
τ^2	0.52 (0.51, 0.52)	0.39 (0.39, 0.40)	0.39 (0.38, 0.40)
ϕ_1	–	0.016 (0.015, 0.016)	0.022 (0.021, 0.023)
ϕ_2	–	–	0.030 (0.029, 0.031)
ν_1	–	0.66 (0.64, 0.67)	0.92 (0.90, 0.93)
ν_2	–	–	0.92 (0.89, 0.93)
p_D	2.94	6526.95	4976.13
DIC	250137	224484.2	222845.1
G	59765.30	42551.08	43117.37
P	59667.15	47603.47	46946.49
D	119432.45	90154.55	90063.86
Time	–	14.53	41.35

The values of the SVC model's goodness of fit metrics suggest that allowing the NDVI regression coefficient to vary spatially improves model fit over that achieved by the SVI model. Figures 2.6(a) and 2.6(b) show maps of posterior estimates for the spatially varying intercept and NDVI, respectively. The clear regional patterns seen in Figure 2.6(b) suggest the relationship between NDVI and biomass does vary spatially—with stronger positive regression coefficients in the Pacific Northwest and northern California areas. Forest in the Pacific Northwest and northern California is dominated by conifers and support the greatest range in biomass per unit area within the entire conterminous United States. The other strong regional pattern seen in Figure 2.6(b) is across western New England, where near zero regression coefficients suggest that NDVI is not as effective at discerning differences in forest biomass. This result is not surprising. For deciduous forests, NDVI can explain variability in low to moderate vegetation



(a) $\beta_0(\mathbf{t})$ (b) $\beta_{NDVI}(\mathbf{t})$

(c) Fitted biomass (d) 95% CI width

Figure 2.6: Forest biomass data analysis using SVC model: (a) Posterior medians of the intercept, (b) NDVI regression coefficients, (c) median of biomass posterior predictive distribution, and (d) range between the upper and lower 95% percentiles of the posterior predictive distribution.

density. However, in high biomass deciduous forests, like those found across western New England, NDVI *saturates* and is no longer sensitive to changes in vegetation structure (Wang et al., 2005). Hence, we see a higher intercept in this region but lower slope coefficient on NDVI. Figures 2.6(c) and 2.6(d) map each location's posterior predictive median and the range between the upper and lower 95% credible interval, respectively, from the SVC model. Figure 2.6(c) shows strong correspondence with the deterministic interpolation of biomass in Figure 2.5(b). The prediction uncertainty in Figure 2.6(d) provides a realistic depiction of the model's ability to quantify forest biomass across the United States.

We also used prediction mean squared error (PMSE) to assess predictive performance. We fit the candidate models using $100,000$ observations and withheld $14,371$ for validation. PMSE for the non-spatial, SVI, and SVC models was 0.52, 0.41, and

0.42 respectively. Lower PMSE for the spatial models, versus the non-spatial model, corroborates the results from the model fit metrics and further supports the need for spatial random effects in the analysis.

2.6 Summary and conclusions

We regard the NNGP as a highly scalable model, rather than a likelihood approximation, for large geostatistical datasets. It significantly outperforms competing low-rank processes such as the GPP, in terms of inferential capabilities as well as scalability. A reference set $\mathcal{S}$ and the resulting neighbor sets (of size m) define the NNGP. Larger m's would increase costs, but there is no apparent benefit to increasing m for larger datasets Selecting $\mathcal{S}$ is akin to choosing the "knots" or "centers" in low-rank methods. While some sensitivity to m and the choice of points in $\mathcal{S}$ is expected, our results indicate that inference is very robust with respect to $\mathcal{S}$ and very modest values of m ($\ll 20$) typically suffice. Larger reference sets may be needed for larger datasets, but its size does not thwart computations. In fact, we observed that a very convenient choice for the reference set is the observed locations.

A potential concern with this choice is that if the observed locations have large gaps, then the resulting NNGP may be a poor approximation of the full Gaussian Process. This arises from the fact that observations at locations outside the reference set are correlated via their respective neighbor sets and large gaps may imply two very near points have very different neighbor sets leading to low correlation. Our simulations indeed reveal that in such a situation, the NNGP covariance field is very flat at points in the gap. However, even with this choice of $\mathcal{S}$ the NNGP model performs at par with the full GP model as the latter also fails to provide strong information about observations located in large gaps. Of course, one can always choose a grid over the entire domain as $\mathcal{S}$ to construct a NNGP with covariance function similar to the full GP. Another choice for $\mathcal{S}$ could be based upon configurations for treed Gaussian processes (Gramacy and Lee, 2008). .

Our simulation experiments revealed that estimation and kriging based on NNGP models closely emulate those from the true Matérn GP models, even for slow decaying covariances . The Matérn covariance function is monotonically decreasing with distance

and satisfies theoretical *screening* conditions, i.e. the ability to predict accurately based on a few neighbors (Stein, 2002). This, perhaps, explains the excellent performance of NNGP models with Matérn covariances. We also investigated the performance of NNGP models using a wave covariance function, which does not satisfy the screening conditions, in a setting where a significant proportion of nearest neighbors had negative correlation with the corresponding locations. The NNGP estimates were still close to the true model parameters and the kriged surface closely resembled the true surface.

Most wave covariance functions (like the damped cosine or the cardinal sine function) produce covariance matrices with several small eigenvalues. The full GP model cannot be implemented for such models because the matrix inversion is numerically unstable. The NNGP model involves much smaller matrix inversions and can be implemented in some cases (e.g. for the damped cosine model). However, for the cardinal sine covariance, the NNGP also faces numerical issues as even the small $m \times m$ covariance matrices are numerically unstable. Bias-adjusted low-rank GPs (Finley et al., 2009) possess a certain advantage in this aspect as the covariance matrix is guaranteed to have eigen values bounded away from zero. However, computations involving low-rank processes with numerically unstable covariance functions cannot be carried out with the efficient Sherman-Woodbury-Morrison type matrix identities and more expensive full Cholesky decompositions will be needed.

Apart from being easily extensible to multivariate and spatiotemporal settings with discretized time, the NNGP can fuel interest in process-based modeling over graphs. Examples include networks, where data arising from nodes are posited to be similar to neighboring nodes. It also offers new modeling avenues and alternatives to the highly pervasive Markov random field models for analyzing regionally aggregated spatial data. Also, there is scope for innovation when space and time are jointly modeled as processes using spatiotemporal covariance functions. One will need to construct neighbor sets both in space and time and effective strategies, in terms of scalability and inference, will need to be explored. Comparisons with alternate approaches (see, e.g., Katzfuss and Cressie, 2012) will also need to be made. Finally, a more comprehensive study on the alternate algorithms, including direct methods for executing sparse Cholesky factorizations, in Section 2.4 is being undertaken. More immediately, we plan to migrate our lower-level `C++` code to the existing `spBayes` package (Finley et al., 2013) in the

`R` statistical environment (`http://cran.r-project.org/web/packages/spBayes`) to facilitate wider user accessibility to NNGP models.

Chapter 3

Non-separable Dynamic Nearest-Neighbor Gaussian Process Models for Large spatio-temporal Data with an Application to Particulate Matter Analysis

3.1 Introduction

Recent years have witnessed considerable growth in statistical modeling of large spatio-temporal datasets; see, for example, the recent books by Gelfand et al. (2010), Cressie and Wikle (2011) and Banerjee et al. (2014) and the references therein for a variety of methods and applications. An especially important domain of application for such models is environmental public health, where analysts and researchers seek map projections for ambient air pollutants measured at monitoring stations and understand the temporal variation in such maps. When inference is sought at the same scale as the observed data, one popular approach is to model the measurements as a time series of

spatial processes. This approach encompasses standard time series models with spatial covariance structures (Pfeifer and Deutsch, 1980a,b; Stoffer, 1986) and dynamic models (Stroud et al., 2001; Gelfand et al., 2005b) among numerous other alternatives.

On the other hand, when inference is sought at arbitrary scales, possibly finer than the observed data (e.g., interpolation over the entire spatial and temporal domains), one constructs stochastic process models to capture dependence using spatio-temporal covariance functions (see, e.g., Cressie and Huang, 1999; Kyriakidis and Journel, 1999; Gneiting, 2002; Stein, 2005; Allcroft and Glasbey, 2003; Gneiting et al., 2007). In modeling ambient air pollution data, it is now customary to meld observed measurements with physical model outputs, where the latter can operate at much finer scales. Inference, therefore, is increasingly being sought at arbitrary resolutions using spatio-temporal process models (see, e.g., Gneiting and Guttorp, 2010) . Henceforth, we focus upon this setting.

While the richness and flexibility of spatio-temporal process models are indisputable, their computational feasibility and implementation pose major challenges for large datasets. Model-based inference usually involves the inverse and determinant of an $n \times n$ spatio-temporal covariance matrix $\mathbf{C}(\boldsymbol{\theta})$, where n is the number of space-time coordinates at which the data have been observed. When $\mathbf{C}(\boldsymbol{\theta})$ has no exploitable structure, matrix computations typically require $\sim n^3$ floating point operations (flops) and storage in the order of n^2 which becomes prohibitive if n is large. Approaches for modeling large covariance matrices in purely spatial settings include low rank models (see, e.g., Higdon, 2001; Kammann and Wand, 2003; Stein, 2007, 2008; Banerjee et al., 2008; Cressie and Johannesson, 2008; Crainiceanu et al., 2008; Rasmussen and Williams, 2005; Finley et al., 2009; Katzfuss, 2016), covariance tapering (see, e.g., Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009; Shaby and Ruppert, 2012; Bevilacqua et al., 2015), approximations using Gaussian Markov Random Fields (GMRF) (see, e.g., Rue and Held, 2005), products of lower dimensional conditional densities (Datta et al., 2016; Vecchia, 1988, 1992; Stein et al., 2004), and composite likelihoods (e.g., Eidsvik et al., 2014). Extensions to spatio-temporal settings include Cressie et al. (2010), Finley et al. (2012) and Katzfuss and Cressie (2012) who extend low-rank spatial processes to dynamic spatio-temporal settings while Xu et al. (2014) who opts for a GMRF approach. All these methods use dynamic models defined on fixed temporal lags and do not lend

themselves easily to continuous spatio-temporal domains.

Spatio-temporal process models for continuous space-time modeling of large datasets have received relatively scant attention. Bai et al. (2012) and Bevilacqua et al. (2012) used composite likelihoods for parameter estimation in a continuous space-time setup. Both these approaches, like their spatial analogues, have focused upon constructing computationally attractive likelihood approximations and have restricted inference only to parameter estimation. Uncertainty estimates are usually based on asymptotic results which are usually inappropriate for irregularly observed datasets. Moreover, prediction at arbitrary locations and time points proceeds by imputing estimates into an interpolator derived from a different process model. This remains expensive for large n and may not reflect predictive uncertainty accurately.

Our current work offers a highly scalable spatio-temporal process for continuous space-time modeling. We expand upon the neighbor-based conditioning set approaches outlined in purely spatial contexts by Vecchia (1988), Stein et al. (2004) and Datta et al. (2016). We derive a scalable version of a spatio-temporal process, which we call the Dynamic Nearest-Neighbor Gaussian Process (DNNGP), using information from smaller sets of neighbors over space and time. This approach offers several benefits. The DNNGP is a well-defined spatio-temporal process whose realizations follow Gaussian distributions with sparse precision matrices. Thus, the DNNGP can act as a sparsity-inducing prior for spatio-temporal random effects in any Bayesian hierarchical model and enables full posterior inference considerably enhancing its applicability. Moreover, it can be used with any spatio-temporal covariance function, thereby accommodating non-separability. Being a process, importantly, allows the DNNGP to provide inference at arbitrary resolutions and, in particular, enables predictions at new spatial locations and time points in posterior predictive fashion. The DNNGP also delivers a substantially superior approximation to the underlying process than, for example, by low rank approximations (see, e.g, Stein, 2014, for problems with low-rank approximations). Finally, storage and memory requirements for a DNNGP model are linear in the number of observations, so it efficiently scales up to massive datasets without sacrificing richness and flexibility in modeling and inference.

The remainder of the article is organized as follows. In Section 3.2 we present the details of a massive environmental pollutants dataset and the need for a full Bayesian analysis. Section 3.3 elucidates a general framework for building scalable spatio-temporal processes and uses it to construct a sparsity-inducing DNNGP over a spatio-temporal domain. Section 3.4 describes efficient schemes for fixed as well as adaptive neighbor selection, which are used in the DNNGP. Section 3.5 details a Bayesian hierarchical model with a DNNGP prior and its implementation using Markov Chain Monte Carlo (MCMC) algorithms. Section 3.6 illustrates the performance of DNNGP using simulated datasets. In Section 3.7 we present a detailed analysis of our environmental pollutants dataset. We conclude the manuscript in Section 6.5 with a brief review and pointers to future research.

3.2 PM_{10} pollution analysis

Exposure to airborne particulate matter (PM) is known to increase human morbidity and mortality (Brunekreef and Holgate, 2002; Loomis et al., 2013; Hoek et al., 2013). In response to these and other health impact studies, regulatory agencies have introduced policies to monitor and regulate PM concentrations. For example, the European Commission's air quality standards limit PM_{10} (PM<10 μm in diameter) concentrations to an average of 50 μg m^{-3} over 24 hours and of 40 μg m^{-3} over a year (European Commission, 2015). Measurements made with standard instruments are considered authoritative, but these observations are sparse and maps at finer scales are needed for monitoring progress with mitigation strategies and for monitoring compliance. Hence, accurately quantifying uncertainty in predicted PM concentrations is critical.

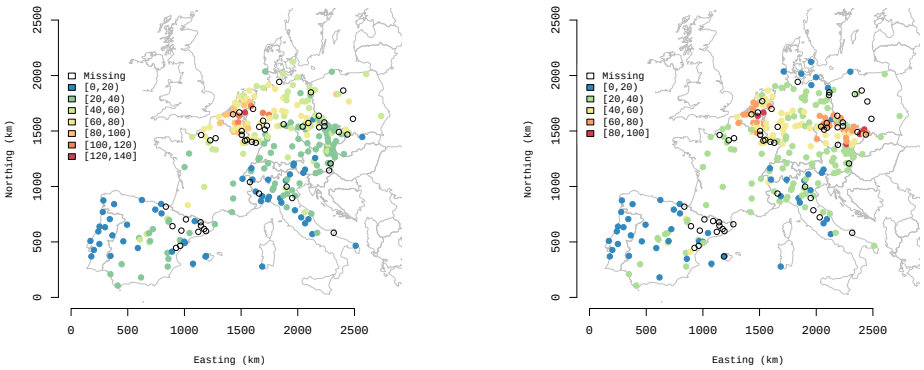
Substantial work has been aimed at developing regional scale chemistry transport models (CTM) for use in generating such maps. CTM's, however, have been shown to systematically underestimate observed PM_{10} concentrations, due to lack of information and understanding about emissions and formation pathways (Stern et al., 2008). Empirical regression (Brauer et al., 2011) or geostatistical models (Lloyd and Atkinson, 2004) are an alternative to CTM's for predicting continuous surfaces of PM_{10}. Empirical models may give accurate results, but are restricted to the conditions under which

they are developed (Manders et al., 2009). Assimilating monitoring station observations and CTM output, with appropriate bias adjustments, has been shown to provide improvements over using either data source alone (van de Kassteele and Stein, 2006; Denby et al., 2008; Candiani et al., 2013; Hamm et al., 2015). In such settings, the CTM output enters as a model covariate and the measured station observations are the response. In addition to delivering more informed and realistic maps, analyses conducted using the models detailed in Section 3.5 can provide estimates of spatial and temporal dependence not accounted for by the CTM and hence provide insights useful for improving the transport models.

We focus on the development and illustration of continuous space-time process models capable of delivering predictive maps and forecasts for PM_{10} and similar pollutants using sparse monitoring networks and CTM output. We coupled observed PM_{10} measurements across central Europe with corresponding output from the LOTOS-EUROS (Schaap et al., 2008) CTM. Inferential objectives included *i*) delivering continuous maps of PM_{10} with associated uncertainty, *ii*) producing statistically valid forecast maps given CTM projections, and *iii*) developing inference on space and time residual structure, i.e., space and time lags, that can help identify lurking processes missing in the CTM. The study area and dataset are the same as those used by Hamm et al. (2015) and the reader is referred to that paper for more background information. Note that the current paper works with a 2-year time series, whereas Hamm et al. (2015) focused on daily analysis of a limited number of pollution events.

3.2.1 Study area

The study domain comprises mainland European countries with a substantial number of available PM_{10} observations. The countries included were Portugal, Spain, Italy, France, Switzerland, Belgium, The Netherlands, Germany, Denmark, Austria, Poland, The Czech Republic, Slovakia and Slovenia. All data were projected to the European Terrestrial Reference System 1989 (ETRS) Lambert Azimuthal Equal-Area (LAEA) projection which gives a coordinate reference system for the whole of Europe.



(a) April 3, 2009 (b) April 5, 2009

Figure 3.1: Observed PM_{10} μg m^{-3} for two example dates.

3.2.2 Observed measurements

Air quality observations for the study area were drawn from the Airbase (*Air* quality data*base*)[1] . Daily PM_{10} concentrations were extracted for January 1 2008 through December 30 2009 resulting in a maximum of M=730 observations at each of $N = 308$ monitoring stations. Airbase daily values are averaged over the within-day hourly values when at least 18 hourly measurements are available, otherwise no data are provided. Airbase monitors are classified by type of area (rural, urban, suburban) and by type (background, industrial, traffic or unknown). Only rural background monitors were used in our study. This is common for comparing measured observations to coarse resolution CTM simulations (Denby et al., 2008). Monitoring stations above 800 m altitude were also excluded. These tend to be located in areas of variable topography and the accuracy of the CTM for locations that shift from inside to outside the atmospheric mixing layer is known to be poor. No further quality control was performed on the data. The locations of the 308 stations used in the subsequent analysis are shown in Figure 3.1 with associated observed and missing PM_{10} for two example dates. Of the 224,840 ($M \times N$) potential observations across 730 day time series and 308 stations, 41,761 observations were missing due to sensor failure or removal, and post-processing removal by Airbase. These missing values were predicted using the proposed models.

[1] http://acm.eionet.europa.eu/databases/airbase (accessed 26 September 2014)

3.2.3 LOTOS-EUROS CTM data

LOTOS-EUROS (v1.8) is a 3D CTM that simulates air pollution in the lower troposphere. The simulator's geographic projection is longitude-latitude at a resolution of 0.50° longitude ×0.25° latitude (approximately 25 km × 25 km). LOTOS-EUROS simulates the evolution of the components of particulate matter separately. Hence, this CTM incorporates the dispersion, formation and removal of sulfate, nitrate, ammonium, sea salt, dust, primary organic and elemental carbon and non-specified primary material, although it does not incorporate secondary organic aerosol. Hendriks et al. (2013) provide a detailed description of LOTOS-EUROS.

The hour-by-hour calculations of European air quality in 2008-2009 were driven by the European Centre for Medium Range Weather Forecasting (ECMWF). Emissions were taken from the MACC (Monitoring Atmospheric Composition and Climate) emissions database (Pouliot et al., 2012). Boundary conditions were taken from the global MACC service (Flemming et al., 2009). The LOTOS-EUROS hourly model output was averaged to daily mean PM_{10} concentrations. LOTOS-EUROS grid cells that were spatially coincident with the Airbase observations were extracted and used as the covariate in the subsequent model.

CTM grid cell values nearest to station locations were used for subsequent model development. No attempt was made to match the spatial support (resolution) of the CTM simulations and station observations. The support of the CTM is 25 km, but the support of the observations is vague. Rural background observations were deliberately chosen because they are distant from urban areas and pollution sources. They are, therefore considered representative of background, ambient pollution conditions and appropriate for matching with moderate resolution CTM-output (Denby et al., 2008; Hamm et al., 2015). This assumption is further backed up by empirical studies indicating that PM_{10} concentrations are dominated by rural background values even in urban areas.

3.3 Scalable Dynamic Nearest-Neighbor Gaussian Processes

Let $\{w(\ell) : \ell \in \mathcal{L}\}$ be a zero-centered continuous spatio-temporal process (see, e.g., Gneiting and Guttorp, 2010, for details), where $\mathcal{L} = \mathcal{S} \times \mathcal{T}$ with $\mathcal{S} \subset \Re^d$ (usually $d = 2$ or 3) is the spatial region, $\mathcal{T} \subset [0, \infty)$ is the time domain and $\ell = (\mathbf{s}, t)$ is a space-time

coordinate with spatial location $\mathbf{s} \in \mathcal{S}$ and time point $t \in \mathcal{T}$. Such processes are specified with a spatio-temporal *covariance function* $\mathrm{Cov}\{w(\ell_i), w(\ell_j)\} = C(\ell_i, \ell_j \,|\, \boldsymbol{\theta})$. For any finite collection $\mathcal{U} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ in $\mathcal{L}$, let $\mathbf{w}_\mathcal{U} = (w(\ell_1)), w(\ell_2), \ldots, w(\ell_n))'$ be the realizations of the process over $\mathcal{U}$. Also, for two finite sets $\mathcal{U}$ and $\mathcal{V}$ containing n and m points in $\mathcal{L}$, respectively, we define the $n \times m$ matrix $\mathbf{C}_{\mathcal{U},\mathcal{V}}(\boldsymbol{\theta}) = \mathrm{Cov}(\mathbf{w}_\mathcal{U}, \mathbf{w}_\mathcal{V} \,|\, \boldsymbol{\theta})$, where the covariances are evaluated using $C(\cdot, \cdot \,|\, \boldsymbol{\theta})$. When $\mathcal{U}$ or $\mathcal{V}$ contains a single point, $\mathbf{C}_{\mathcal{U},\mathcal{V}}$ is a row or column vector. A valid spatio-temporal covariance function ensures that $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ is positive definite for any finite set $\mathcal{U}$. In particular, for spatio-temporal Gaussian processes, $\mathbf{w}_\mathcal{U}$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta}))$ and the (i,j)th element of $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ is $C(\ell_i, \ell_j \,|\, \boldsymbol{\theta})$.

Storage and computations involving $\mathbf{C}_{\mathcal{U},\mathcal{U}}(\boldsymbol{\theta})$ can become impractical when n is large relative to the resources available. For full Bayesian inference on a continuous domain, we seek a scalable (in terms of flops and storage) spatio-temporal Gaussian process that will provide an excellent approximation to a full spatio-temporal process with any specified covariance function. We outline a general framework that first uses a set of points in $\mathcal{L}$ to construct a computationally efficient approximation for the random field and extends the finite dimensional distribution over this set to a process. To ease the notation, we will suppress the explicit dependence of matrices and vectors on $\boldsymbol{\theta}$ whenever the context is clear.

Let $\mathcal{R} = \{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$ be a fixed finite set of r points in $\mathcal{L}$. We refer to $\mathcal{R}$ as a *reference set*. We construct a spatio-temporal process $w(\ell)$ on $\mathcal{L}$ by first specifying $\mathbf{w}_\mathcal{R} = (w(\ell_1^*), w(\ell_2^*), \ldots, w(\ell_r^*))' \sim N(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))$, where $\mathbf{K}(\boldsymbol{\theta})$ is any $r \times r$ positive definite matrix and then defining

$$w(\ell) = \sum_{i=1}^{r} a_i(\ell) w(\ell_i^*) + \eta(\ell) \text{ for any } \ell \notin \mathcal{R} \,, \tag{3.1}$$

where $\eta(\ell)$ is a zero-centered Gaussian process independent of $\mathbf{w}_\mathcal{R}$ such that for any two distinct points in $\mathcal{L}$, $\mathrm{Cov}\{\eta(\ell_i), \eta(\ell_j)\} = 0$.

Observe that $w(\ell)$ in (3.1) is a well defined spatio-temporal Gaussian process on $\mathcal{L}$ for *any* choice of $a_i(\ell)$'s, as long as $\mathbf{K}(\boldsymbol{\theta})$ is positive definite. For example, $w(\ell)$ is a Gaussian process with covariance function $C(\cdot, \cdot \,|\, \boldsymbol{\theta})$ if we set $\mathbf{K}(\boldsymbol{\theta}) = \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})$, $\mathbf{a}(\ell) = \mathbf{C}_{\mathcal{R},\mathcal{R}}^{-1}\mathbf{C}_{\mathcal{R},\ell}$ where $\mathbf{a}(\ell)$ is $r \times 1$ with elements $a_i(\ell)$, and $\eta(\ell) \overset{ind}{\sim} N\left(0, C(\ell, \ell \,|\, \boldsymbol{\theta}) - \mathbf{C}_{\ell,\mathcal{R}}\mathbf{C}_{\mathcal{R},\mathcal{R}}^{-1}\mathbf{C}_{\mathcal{R},\ell}\right)$. Now (3.1) represents the 'kriging' equation for a location ℓ based on observations over $\mathcal{R}$

(Cressie and Wikle, 2011). Dimension reduction can be achieved with suitable choices for $\mathbf{K}(\boldsymbol{\theta})$ and $\mathbf{a}(\ell)$. Low-rank spatio-temporal processes emerge when we choose $\mathcal{R}$ to be a smaller set of 'knots' (or 'centers'). Additionally, specifying $\eta(\ell)$ to be a diagonal or sparse residual process yields $w(\ell)$ to be a non-degenerate (or bias-adjusted) low rank Gaussian Process (Banerjee et al., 2008; Finley et al., 2009; Sang and Huang, 2012).

Because of demonstrably impaired inferential performance of low-rank models in purely spatial contexts at scales similar to ours (see, e.g., Stein, 2014; Datta et al., 2016), we use the framework in (3.1) to construct a class of sparse spatio-temporal process models. To be specific, let the reference set $\mathcal{R}$ be an enumeration of $r = MN$ points in $\mathcal{L}$, so that each ℓ_i^* in $\mathcal{R}$ corresponds to some $(\mathbf{s}_j, t_k)$ for $j = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, M$. For any $\ell_i^* = (\mathbf{s}_j, t_k)$ in $\mathcal{R}$ we define a *history set* $H(\ell_i^*)$ as the collection of all locations observed at times before t_k and of all points at time t_k with spatial locations in $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{j-1}\}$. Thus, $H(\ell_i^*) = \{(\mathbf{s}_p, t_q) \,|\, p = 1, 2, \ldots, N,\, q = 1, 2, \ldots, (k-1)\} \cup \{(\mathbf{s}_p, t_k) \,|\, p = 1, 2, \ldots, (j-1)\}$. For any location ℓ_i^* in $\mathcal{R}$, let $N(\ell_i^*)$ be a subset of the history set $H(\ell_i^*)$. Also, for any location $\ell \notin \mathcal{R}$, let $N(\ell)$ denote any finite subset of $\mathcal{R}$. We refer to the sets $N(\ell)$ as a 'neighbor set' for the location ℓ and describe their construction later.

We now turn to our choices for $\mathbf{K}(\boldsymbol{\theta})$ and $\mathbf{a}(\ell)$ in (3.1). Let $w(\ell) \sim GP(0, C(\cdot, \cdot \,|\, \boldsymbol{\theta}))$. We choose $\mathbf{K}(\boldsymbol{\theta})$ to effectuate a sparse approximation for the joint density of the realizations of $w(\ell)$ over $\mathcal{R}$, i.e., $N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta}))$. Adapting the ideas underlying likelihood approximations in Vecchia (1988) and Datta et al. (2016), we specify $\mathbf{K}(\boldsymbol{\theta})$ to be the $r \times r$ matrix such that

$$\begin{aligned} N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{C}_{\mathcal{R},\mathcal{R}}(\boldsymbol{\theta})) &= \prod_{i=1}^{r} p(w(\ell_i^*) \,|\, \mathbf{w}_{H(\ell_i^*)}) \\ &\approx \prod_{i=1}^{r} p(w(\ell_i^*) \,|\, \mathbf{w}_{N(\ell_i^*)}) = N(\mathbf{w}_{\mathcal{R}} \,|\, \mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))\,. \end{aligned} \tag{3.2}$$

Here, $H(\ell_1^*)$ is the empty set (hence, so is $N(\ell_1^*)$) and we define $p(w(\ell_1^*) \,|\, \mathbf{w}_{H(\ell_1^*)}) = p(w(\ell_1^*) \,|\, \mathbf{w}_{N(\ell_1^*)}) = p(w(\ell_1^*))$. The underlying idea behind the approximation in Equation 3.2 is to compress the conditioning sets from $H(\ell_i^*)$ to $N(\ell_i^*)$ so that the resulting approximation is a multivariate normal distribution with a sparse precision matrix $\mathbf{K}^{-1}$. This implies

$$E[w(\ell_i^*) \,|\, \mathbf{w}_{H_{\ell_i^*}}] = \mathrm{E}[w(\ell_i^*) \,|\, \mathbf{w}_{N(\ell_i^*)}] = \mathbf{a}'_{N(\ell_i^*)} \mathbf{w}_{N(\ell_i^*)} \tag{3.3}$$

where $\mathbf{a}_{N(\ell_i^*)} = \mathbf{C}^{-1}_{N(\ell_i^*),N(\ell_i^*)}\mathbf{C}_{N(\ell_i^*),\ell_i^*}$. Also, $\mathbf{K}$ is determined by $\mathbf{C}_{\mathcal{R},\mathcal{R}}$ because $\mathbf{K}^{-1} = \mathbf{V}'\mathbf{F}^{-1}\mathbf{V}$, where $\mathbf{F}$ is a diagonal matrix with diagonal entries $f_{\ell_i^*} = \mathrm{Var}(w(\ell_i^*) \mid \mathbf{w}_{N(\ell_i^*)}) = C(\ell_i^*, \ell_i^* \mid \boldsymbol{\theta}) - \mathbf{C}_{\ell_i^*,N(\ell_i^*)}\mathbf{C}^{-1}_{N(\ell_i^*),N(\ell_i^*)}\mathbf{C}_{N(\ell_i^*),\ell_i^*}$ and $\mathbf{V}$ is the $r \times r$ matrix with entries $v_{i,j}$ such that $v_{i,i} = 1$ and $v_{i,j} = 0$ whenever $\ell_i^* \notin N(\ell_j^*)$. The remaining entries in column j of $\mathbf{V}$ are specified by setting the subvector $\mathbf{V}_{c(\ell_j^*),j} = -\mathbf{a}_{N(\ell_j^*)}$, where $c(\ell_j^*) = \{i \mid \ell_i^* \in N(\ell_j^*)\}$. If $m(<< r)$ denotes the limiting size of the neighbor sets $N(\ell)$, then the columns of $\mathbf{V}$ are sparse with at most $m+1$ non-zero elements. Consequently, $\mathbf{K}^{-1}$ has at most $O(rm^2)$ non-zero elements (this is the spatial-temporal analogue of the result in Datta et al., 2016). Hence, the approximation in (3.2) produces a sparsity-inducing proper prior distribution for the spatio-temporal random effects over $\mathcal{R}$ that closely approximates the realizations from a $GP(0, C(\cdot,\cdot \mid \boldsymbol{\theta}))$.

Turning to the vector of coefficients $\mathbf{a}(\ell)$ in (3.1), we extend the idea in (3.3) to any point $\ell \notin \mathcal{R}$ by requiring that $\mathrm{E}[w(\ell) \mid \mathbf{w}_{\mathcal{R}}] = \mathrm{E}[w(\ell) \mid \mathbf{w}_{N(\ell)}]$. This is achieved by setting $a_i(\ell) = 0$ in (3.1) whenever $\ell_i^* \notin N(\ell)$ for any point $\ell \notin \mathcal{R}$. Hence, if $N(\ell)$ contains m points, then at most m of the elements in the $r \times 1$ vector $\mathbf{a}(\ell)$ can be nonzero. These nonzero entries are determined from the above conditional expectation given $N(\ell)$. To be precise, if $\mathbf{a}_{N(\ell)}$ is the $m \times 1$ vector of these m entries, then we solve $\mathbf{C}_{N(\ell),N(\ell)}\mathbf{a}_{N(\ell)} = \mathbf{C}_{N(\ell),\ell}$ for $\mathbf{a}_{N(\ell)}$. Also note that $\mathbf{a}'(\ell)\mathbf{w}_{\mathcal{R}} = \mathbf{a}'_{N(\ell)}\mathbf{w}_{N(\ell)}$. Finally, to complete the process specifications in (3.1), we specify $\eta(\ell) \overset{ind}{\sim} N(0, f_\ell)$, where $f_\ell = \mathrm{Var}(w(\ell) \mid \mathbf{w}_{N(\ell)}) = C(\ell, \ell \mid \boldsymbol{\theta}) - \mathbf{C}_{\ell,N(\ell)}\mathbf{C}^{-1}_{N(\ell),N(\ell)}\mathbf{C}_{N(\ell),\ell}$. The covariance function $\tilde{C}(\cdot,\cdot \mid \boldsymbol{\theta})$ of the resulting Gaussian Process is given by:

$$\tilde{C}(\ell_i, \ell_j \mid \boldsymbol{\theta}) = \begin{cases} K_{p,q} \text{ if } \ell_i = \ell_p^* \text{ and } \ell_j = \ell_q^* \text{ are both in } \mathcal{R} \\ \mathbf{a}'(\ell_i)\mathbf{K}_{*q} \text{ if } \ell_i \notin \mathcal{R} \text{ and } \ell_j = \ell_q^* \in \mathcal{R} \\ \mathbf{a}'(\ell_i)\mathbf{K}\mathbf{a}(\ell_j) + I(\ell_i = \ell_j) f_{\ell_i} \text{ if } \ell_i \notin \mathcal{R} \text{ and } \ell_j \notin \mathcal{R}\,, \end{cases} \tag{3.4}$$

where $K_{p,q}$ is element (p,q) and $\mathbf{K}_{*q}$ is column q in $\mathbf{K}$.

Owing to the sparsity of $\mathbf{K}^{-1}$, the likelihood $N(\mathbf{w}_{\mathcal{R}} \mid \mathbf{0}, \mathbf{K})$ can be evaluated using $O(rm^3)$ flops for any given $\boldsymbol{\theta}$. Substantial computational savings accrue because m is usually very small (also see later sections). Furthermore as $\eta(\ell)$ yields a diagonal covariance matrix and $\mathbf{a}(\ell)$ has at most m non-zero elements, for any finite set $\mathcal{V}$ outside $\mathcal{R}$, the flop count for computing the density $p(\mathbf{w}_{\mathcal{V}} \mid \mathbf{w}_{\mathcal{R}}, \boldsymbol{\theta})$ will be linear in the size of $\mathcal{V}$. We have now constructed a scalable spatio-temporal Gaussian Process from a *parent* spatio-temporal $GP(0, C(\cdot,\cdot \mid \boldsymbol{\theta}))$ using small neighbor sets $N(\ell)$. We denote

this *Dynamic Nearest Neighbor Gaussian Process* (DNNGP) as $DNNGP(0, \tilde{C}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$, where $\tilde{C}(\cdot,\cdot \,|\, \boldsymbol{\theta}))$ denotes the covariance function of this new GP.

3.4 Constructing Neighbor-Sets

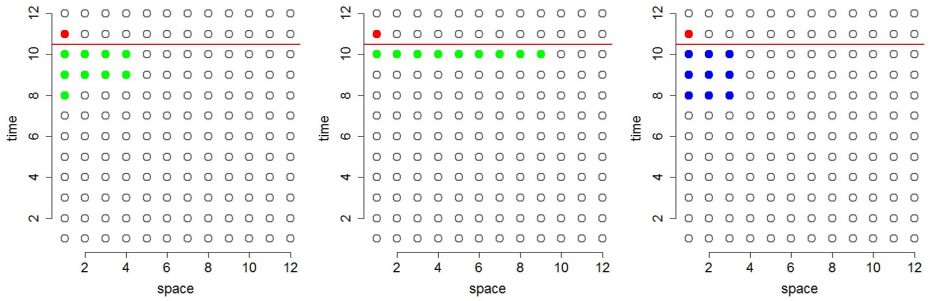
3.4.1 Simple Neighbor Selection

Spatial correlation functions usually decay with increasing inter-site distance, so the set of nearest neighbors based on the inter-site distances represents locations exhibiting highest correlation with the given location. This has motivated use of nearest neighbors to construct these small neighbor sets (Vecchia, 1988; Datta et al., 2016). On the other hand, spatio-temporal covariances between two points typically depend on the spatial as well as the temporal lag between the points. To be specific, non-separable isotropic spatio-temporal covariance functions can be written as $C((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2) \,|\, \boldsymbol{\theta}) = C(h,u \,|\, \boldsymbol{\theta})$ where $h = ||\mathbf{s}_1 - \mathbf{s}_2||$ and $u = |t_1 - t_2|$. This often precludes defining any universal distance function $d : (\mathcal{S} \times \mathcal{T})^2 \to \mathbb{R}^+$ such that $C((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2) \,|\, \boldsymbol{\theta})$ will be monotonic with respect to $d((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2))$ for all choices of $\boldsymbol{\theta}$.

In the light of the above discussion, we define "nearest neighbors" in a spatio-temporal domain using the spatio-temporal covariance function itself as a proxy for distance. To elucidate, for any three points $(\mathbf{s}_1,t_1)$, $(\mathbf{s}_2,t_2)$ and $(\mathbf{s}_3,t_3)$, we say that $(\mathbf{s}_1,t_1)$ is nearer to $(\mathbf{s}_2,t_2)$ than to $(\mathbf{s}_3,t_3)$ if $C((\mathbf{s}_1,t_1),(\mathbf{s}_2,t_2) \,|\, \boldsymbol{\theta}) \;\; > C((\mathbf{s}_1,t_1),(\mathbf{s}_3,t_3) \,|\, \boldsymbol{\theta})$. Subsequently, this definition of "distance" is used to find m nearest neighbors for any location.

Of course, this choice of nearest neighbors depends on the choice of the covariance function C and $\boldsymbol{\theta}$. Since the purpose of the DNNGP is to provide a scalable approximation of the parent GP, we always choose $C(\cdot,\cdot \,|\, \boldsymbol{\theta})$ to be same as the covariance function of the parent GP. However, for every location $(\mathbf{s}_i,t_j)$, its neighbor set, denoted by $N_\theta(\mathbf{s}_i,t_j)$, still depends on $\boldsymbol{\theta}$. This is illustrated in Figures 3.2(a) and 3.2(b) which shows how neighbor sets can differ drastically based on the choice of $\boldsymbol{\theta}$.

In most applications, $\boldsymbol{\theta}$ is unknown precluding the use of these newly defined neighbor sets $N_\theta(\mathbf{s}_i,t_j)$ to construct the DNNGP. We propose a simple intuitive method to construct neighbor sets. We choose m to be a perfect square and construct a



(a) $N_{\theta=1}(s_i, t_j)$ (b) $N_{\theta=2}(s_i, t_j)$ (c) Simple neighbor sets

Figure 3.2: True and simple neighbor sets for a 12×12 spatio-temporal dataset with one-dimensional spatial domain and covariance function $C((s_1, t_1), (s_2, t_2) \,|\, \theta) = \exp(-|s_1 - s_2|^2 - \theta|t_1 - t_2|^2)$. All points below the red horizontal line constitute the history set for the red point (s_i, t_j). Green points denote $N_\theta(s_i, t_j)$ – the sets of $m(= 9)$ true nearest neighbors with $\theta = 1$ (figure (a)) and $\theta = 2$ (figure (b)). The blue points in figure (c) denotes the simple neighbor set.

simple neighbor set of size m using $\sqrt{m}$ spatial nearest neighbors and $\sqrt{m}$ temporal nearest neighbors. Figure 3.2(c) illustrates the simple neighbor set of size $m = 9$ for the red point. In order to formally define the simple neighbor sets, we denote $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$, $S_i = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ and $T = \{t_1, t_2, \ldots, t_M\}$. Furthermore, for any finite set of spatial locations $V \subseteq S$, let $A(\mathbf{s}, V, m)$ denote the set of m nearest neighbors in V for the location $\mathbf{s}$. For any point $(\mathbf{s}_i, t_j) \in \mathcal{R}$ we define the simple neighbor sets

$$N(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{\sqrt{m}-1} \{(\mathbf{s}, t_{j-k}) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S, \sqrt{m})\} \bigcup \{(\mathbf{s}, t_j) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S_i, \sqrt{m})\} \tag{3.5}$$

The above construction implies that the neighbor set for any point in $\mathcal{R}$ consists of $\sqrt{m}$ spatial nearest neighbors of the preceding $\sqrt{m}$ time points. For arbitrary $(\mathbf{s}, t) \notin \mathcal{R}$, $N(\mathbf{s}, t)$ is simply defined as the Cartesian product of $\sqrt{m}$ nearest neighbors for $\mathbf{s}$ in $\mathcal{S}$ with $\sqrt{m}$ nearest neighbors of t in $\mathcal{T}$.

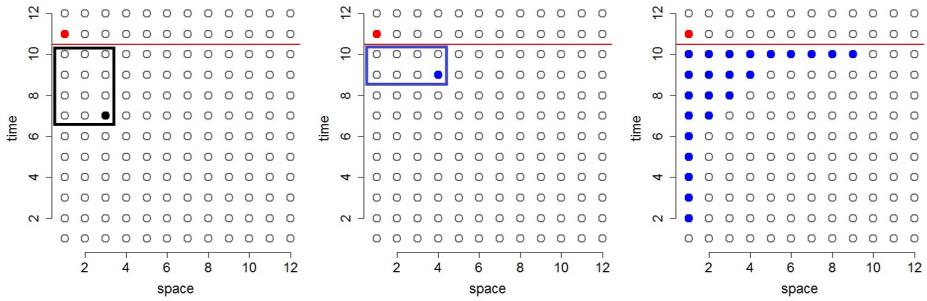
In many applications, one desirable property of the spatio-temporal covariance functions is *natural monotonicity*, i.e. $C(h, u)$ is decreasing in h for fixed u and decreasing

in u for fixed h. All Matèrn-based space-time separable covariances and many non-separable classes of covariance functions possess this property (Stein, 2013; Omidi and Mohammadzadeh, 2015). If $C(\cdot,\cdot \mid \boldsymbol{\theta})$ possesses natural monotonicity, then $N(\mathbf{s}_i, t_j)$ defined in Equation 3.5 is guaranteed to contain at least $\sqrt{m}-1$ nearest neighbors of $(\mathbf{s}_i, t_j)$ in $H(\mathbf{s}_i, t_j)$. Thus, the neighbor sets defined above do not depend on any parameter and, for any value of $\boldsymbol{\theta}$, will contain a few nearest neighbors.

3.4.2 Adaptive Neighbor Selection

The simple neighbor selection scheme described in Section 3.4.1 does not depend on $\boldsymbol{\theta}$ and is undoubtedly useful for fast implementation of the DNNGP. However, for some values of $\boldsymbol{\theta}$, the neighbor sets may often consist of very few nearest neighbors. This issue is illustrated in Figure 3.2 where the simple neighbor set (blue points) contained 7 out of 9 true nearest neighbors (green points) for $\theta = 1$ but only 3 out of 9 true nearest neighbors for $\theta = 2$. We see that for different choices of the covariance parameters the simple neighbor sets contain different proportions of the true nearest neighbors. The problem is exacerbated in extreme cases with variation only along the spatial or temporal direction. In such cases, the neighbor sets defined in (3.5) will contain only about $\sqrt{m}$ nearest neighbors and $m - \sqrt{m}$ uncorrelated points.

Ideally, if $\boldsymbol{\theta}$ was known, one could have simply evaluated the pairwise correlations between any point $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$ and all points in its history set $H(\mathbf{s}_i, t_j)$ to obtain $N_\theta(\mathbf{s}_i, t_j)$ — the set of m true nearest neighbors. In practice, however, we encounter a computational roadblock because $\boldsymbol{\theta}$ is unknown and for every new value of $\boldsymbol{\theta}$ in an iterative optimizer or Markov Chain Monte Carlo sampler, we need to redo the search for the neighbor sets within the history sets. As the history sets are typically large this is computationally challenging. For example, in Figure 3.2, the history set for the red point is composed of all points below the red horizontal line. So, evaluating the pairwise correlations required for updating neighbor sets of all points in $\mathcal{R}$ and n datapoints outside $\mathcal{R}$, will use $O(r^2 + nr)$ flops at each iteration. The reference set $\mathcal{R}$ is typically chosen to match the scale of the observed dataset to achieve a reasonable approximation of the parent GP by DNNGP. Hence, for large datasets this updating becomes a deterrent. In fact, Vecchia (1988) and Stein et al. (2004) admit that this challenge has inhibited the use of correlation based neighbor sets in a spatial setting.



(a) Ineligible point (b) Eligible point (c) Full eligible set

Figure 3.3: Construction of eligible sets for finding nearest neighbor sets of size $m = 9$: In figure (a) the black point is ineligible bacause the black rectangle contains more than $m = 9$ points. In figure (b) the blue point will belong to $E(\mathbf{s}_i, t_j)$ as the blue rectangle contains less than $m = 9$ points. Figure (c) shows the final eligible set obtained by repeating this algorithm for all points in the history set (below the red line).

Jones and Zhang (1997) permitted locations within a small prefixed temporal lag of a given location to be eligible for neighbors. However, this assumption will fail to capture any long term temporal dependence present in the datasets.

We now provide an algorithm that efficiently updates the neighbor sets after every update of $\boldsymbol{\theta}$. The underlying idea is to restrict the search for the neighbor sets to carefully constructed small subsets of the history sets. These small *eligible sets* $E(\mathbf{s}_i, t_j)$ are constructed in such a manner that, despite being much smaller than the history sets, they are guaranteed to contain the true nearest neighbor sets $N_\theta(\mathbf{s}_i, t_j)$ for all choices of the parameter $\boldsymbol{\theta}$. So, for each $\boldsymbol{\theta}$ we can evaluate the pairwise correlations between $(\mathbf{s}_i, t_j)$ and only the points in $E(\mathbf{s}_i, t_j)$ and still find the true set of m-nearest neighbors.

Figures 3.3(a) and 3.3(b) illustrate how to determine which points belong to $E(\mathbf{s}_i, t_j)$. Let h and u denote the spatial and temporal lags with the black point and the red point in Figure 3.3(a). All other points in the black rectangle have spatial lag $\leq h$ and temporal lag $\leq u$ with the red point. So if the covariance function $C(h, u \,|\, \boldsymbol{\theta})$ possess natural monotonicity, the black point has lowest correlation with the red point among all the points in the black rectangle. For the black point to be in the the set of m nearest neighbors $N_\theta(\mathbf{s}_i, t_j)$ for any $\boldsymbol{\theta}$, all other points in the black rectangle should also

be included. Since, this is not possible as the black rectangle contains 12 points and $m = 9$, the black point becomes ineligible. By a similar logic, the blue rectangle in Figure 3.3(b) contains $8(< m)$ points and is included in $E(\mathbf{s}_i, t_j)$. Proceeding like this, we can easily determine the entire eligible set (Figure 3.3(c)) without any knowledge of the parameter $\boldsymbol{\theta}$.

We now provide a formal construction of the eligible sets. Recall from Section 3.4.1 that, for any location s, $A(\mathbf{s}, V, m)$ is the set of m-nearest neighbors of $\mathbf{s}$ in V. So $\mathbf{s} \in V$ implies that $\mathbf{s} \in A(\mathbf{s}, V, m)$ for all $m \geq 1$. For each $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$, we define the *eligible set*

$$E(\mathbf{s}_i, t_j) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t_{j-k}) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S, [m/k])\} \bigcup \{(\mathbf{s}, t_j) \,|\, \mathbf{s} \in A(\mathbf{s}_i, S_i, m)\} \tag{3.6}$$

where for any positive number x, $[x]$ denotes the greatest integer not exceeding x. So the eligible set for a space-time point consists of m-nearest neighbors from the time levels j and $j-1$, $[m/2]$ nearest neighbors from time level $j-2$ and so on upto $[m/m] = 1$ nearest neighbor from time level $j-m$. This is also illustrated in Figure 3.3(c). So the size of $E(\mathbf{s}_i, t_j)$ does not exceed $m + \sum_{k=1}^{m} [m/k]$. As m is typically chosen to be around 20, this sum is approximately $4m$.

For any point t outside T, let $t[k]$ denote the k^{th} nearest time point of t in T. Then, we define the eligible set for any $(\mathbf{s}, t)$ outside $\mathcal{R}$ as

$$E(\mathbf{s}, t) = \bigcup_{k=1}^{m} \{(\mathbf{s}, t[k]) \,|\, \mathbf{s} \in A(\mathbf{s}, S, [m/k])\} \tag{3.7}$$

The eligible sets do not depend on the covariance parameters $\boldsymbol{\theta}$. We now show that for any point $(\mathbf{s}, t)$ in $\mathcal{L}$, the eligible set $E(\mathbf{s}, t)$ defined by Equations 3.6 and 3.7 contains m-nearest neighbors of $(\mathbf{s}, t)$ for all values of $\boldsymbol{\theta}$ as long as the underlying covariance function $C(h, u \,|\, \boldsymbol{\theta})$ possess natural monotonicity.

Proposition 1. *If $C(h, u \,|\, \boldsymbol{\theta})$ satisfies natural monotonicity defined in Section 3.4.1 for every value of $\boldsymbol{\theta}$, then, for every $(\boldsymbol{s}, t)$, the eligible set $E(\boldsymbol{s}, t)$ defined in Equations 3.6 and 3.7 contains $N_\theta(\boldsymbol{s}, t)$ for all $\boldsymbol{\theta}$*

Proof. We only prove for $(\mathbf{s}, t) = (\mathbf{s}_i, t_j) \in \mathcal{R}$. The proof for $(\mathbf{s}, t) \notin \mathcal{R}$ is similar. We assume that $(\mathbf{s}_u, t_{j-k}) \in N_\theta(\mathbf{s}_i, t_j)$ for some $\boldsymbol{\theta}$, $u \leq N$ and $k \geq 1$. Also let $\mathbf{s}_i[l]$

denote the l^{th} nearest neighbor of $\mathbf{s}_i$ among $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$. So, $\mathbf{s}_u = \mathbf{s}_i[l]$ for some $l \geq 1$. Therefore, by natural monotonicity of C, we have $C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-k}) \,|\, \boldsymbol{\theta}) \geq C((\mathbf{s}_i, t_j), (\mathbf{s}_i[l], t_{j-k}) \,|\, \boldsymbol{\theta})$ for all $1 \leq a \leq l$. One more application of natural monotonicity implies that $C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-b}) \,|\, \boldsymbol{\theta}) > C((\mathbf{s}_i, t_j), (\mathbf{s}_i[a], t_{j-k}) \,|\, \boldsymbol{\theta})$ for all $1 \leq b \leq k$. As $(\mathbf{s}_u, t_{j-k}) \in N_\theta(\mathbf{s}_i, t_j)$, then so does $(\mathbf{s}_i[a], t_{j-b})$ for all $a \leq l$ and $b \leq k$. Therefore, $lk \leq m$ i.e. $l \leq [m/k]$. □

Proposition 1 proves that eligible sets are guaranteed to contain the neighbor sets for all choices of $\boldsymbol{\theta}$. This result has substantial consequences because the size of the eligible sets is approximately equal to $4m$. The eligible sets needs to be calculated only once before the MCMC as they are free of any parameter choices. Subsequently, for every new update of $\boldsymbol{\theta}$ in a MCMC sampler or an iterative solver, one can search for a new set of m-nearest neighbors $N_\theta(\mathbf{s}_i, t_j)$ only within the eligible sets and use $N_\theta(\mathbf{s}_i, t_j)$ as the conditioning sets to construct the DNNGP. We summarize the MCMC steps of the DNNGP with adaptive neighbor selection in Algorithm 1.

Algorithm 1 Algorithm for adaptive neighbor selection in dynamic NNGP

1: Compute the eligible sets $E(\mathbf{s}_i, t_j)$ for all $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$ from Eqn. (3.6)

2: At the l^{th} iteration of the MCMC:

   (a) Calculate $C((\mathbf{s}, t), (\mathbf{s}_i, t_j) \,|\, \boldsymbol{\theta}^{(l)})$ for all $(\mathbf{s}, t)$ in $E(\mathbf{s}_i, t_j)$

   (b) Define $N_\theta(\mathbf{s}_i, t_j)^{(l)}$ as the set of m locations in $E(\mathbf{s}_i, t_j)$ which maximizes $C((\mathbf{s}, t), (\mathbf{s}_i, t_j) \,|\, \boldsymbol{\theta}^{(l)})$

   (c) Repeat steps (a) and (b) for all $(\mathbf{s}_i, t_j)$ in $\mathcal{R}$

   (d) Update $\boldsymbol{\theta}^{(l+1)}$ based on the new set of neighbor sets computed in step (c) using Metropolis step specified in (3.12)

3: Repeat Step 2 for N MCMC iterations

As the size of the sets are approximately $4m$, for every $(\mathbf{s}_i, t_j)$ we need to evaluate only $4m$ pairwise correlations. So the total computational complexity of the search is now reduced to $O(4m(n + r))$ from $O(nr + r^2)$. This is at par with the scale of implementing the remainder of the algorithm. With this adaptive neighbor selection scheme we gain the advantage of selecting the set of m-nearest neighbors at every

update while retaining the scalability of the DNNGP. Parallel computing resources, if available, can be greatly utilized to further reduce computations as the search for eligible sets for each point (Algorithm 1: Step (c)) can proceed independent of one another.

3.5 Bayesian DNNGP model

We consider a spatio-temporal dataset observed at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$ and at time points $t_1, t_2, \ldots, t_M$. Note that there may not be data for all locations at all time points i.e. we allow missing data. Let $\{\ell_1, \ell_2, \ldots, \ell_n\}$ be an enumeration of $n = MN$ points in $\mathcal{L}$, where each ℓ_i is an ordered pair $(\mathbf{s}_j, t_k)$. Let $y(\ell_i)$ be a univariate response corresponding to ℓ_i and let $\mathbf{x}(\ell_i)$ be a corresponding $p \times 1$ vector of spatio-temporally referenced predictors. A spatio-temporal regression model relates the response and the predictors as

$$y(\ell_i) = \mathbf{x}'(\ell_i)\boldsymbol{\beta} + w(\ell_i) + \epsilon(\ell_i) \ , \quad i = 1, 2, \ldots, MN \ , \tag{3.8}$$

where $\boldsymbol{\beta}$ denotes the coefficient vector for the predictors, $w(\ell_i)$ is the spatio-temporally varying intercept and $\epsilon(\ell_i)$ is the random noise customarily assumed to be independent and identically distributed copies from $N(0, \tau^2)$.

Usually $w(\ell_i)$'s are modeled as realizations of a spatio-temporal GP. To ensure scalability, we will construct a DNNGP from a parent GP with a non-separable spatio-temporal isotropic covariance function $C((\mathbf{s}+\mathbf{h}, t+u), (\mathbf{s}, t) \,|\, \boldsymbol{\theta})$, introduced by Gneiting (2002),

$$\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)(a|u|^{2\alpha}+1)^{\delta+\kappa}} \times \left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\kappa/2}}\right)^{\nu} \times K_\nu\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\kappa/2}}\right) \ , \tag{3.9}$$

where $\mathbf{h}$ and u refers to the spatial and temporal lags between $(\mathbf{s}+\mathbf{h}, t+u)$ and $(\mathbf{s}, t)$ and $\boldsymbol{\theta} = \{\sigma^2, \alpha, \kappa, \delta, \nu, a, c\}$. The spatial covariance function at temporal lag zero corresponds to the Whittle-Matern class with marginal variance σ^2, smoothness parameter ν and decay parameter c. The parameters α and a control smoothness and decay, respectively, for the temporal process, while κ captures non-separability between space and time.

A straightforward choice of the reference set $\mathcal{R}$ is the set $\{\ell_1, \ell_2, \ldots, \ell_n\}$. While this set will typically be large, its size does not adversely affect the computations. This choice has been shown to yield excellent approximations to the parent random field

(Vecchia, 1988; Stein et al., 2004). Also, while several alternate choices of reference sets (like choosing the points over a regular grid) are possible, it is unlikely they will provide any additional computational or inferential benefits; this has been demonstrated in purely spatial contexts by Datta et al. (2016). Hence, we choose $\mathcal{R} = \{\ell_1, \ell_2, \ldots, \ell_n\}$, i.e., $\ell_i^* = \ell_i$ for $i = 1, 2, \ldots, n$.

A full hierarchical model with a DNNGP prior on $w(\ell)$ is given by

$$p(\boldsymbol{\theta}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \mathbf{V}_\beta) \times N(\mathbf{w}_\mathcal{R} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_{\mathcal{R},\mathcal{R}}) \\ \times \prod_{i=1}^{n} N(\mathbf{y}(\ell_i) \,|\, \mathbf{x}(\ell_i)'\boldsymbol{\beta} + \mathbf{w}(\ell_i), \tau^2) \,, \tag{3.10}$$

where $p(\boldsymbol{\theta})$ is the prior on $\boldsymbol{\theta}$, and $IG(\tau^2 \,|\, a_\tau, b_\tau)$ denotes the Inverse-Gamma density with shape a_τ and rate b_τ. Below we describe an efficient MCMC algorithm using Gibbs and Metropolis steps only to carry out full inference from the posterior in Equation 4.3.

3.5.1 Gibbs' sampler steps

Let S_o be the points in $\mathcal{R}$ where the $y(\ell_i)$'s is observed and $I(\ell_i)$ denote the binary indicator for presence or absence of data at ℓ_i. Let $\mathbf{y}$ be the $n_o \times 1$ vector formed by stacking the responses observed and $\mathbf{X}$ be the corresponding $n_o \times p$ design matrix. The full conditional distribution of $\boldsymbol{\beta}$ is $N(\mathbf{V}_\beta^* \boldsymbol{\mu}_\beta^*, \mathbf{V}_{\boldsymbol{\beta}}^*)$ where $\mathbf{V}_\beta^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}'\mathbf{X}/\tau^2)^{-1}$ and $\boldsymbol{\mu}_\beta^* = (\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'(\mathbf{y} - \mathbf{w}_{S_o})/\tau^2)$. The full conditional distribution of τ^2 follows $IG\left(a_\tau + \frac{n_o}{2}, b_\tau + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{w}_{S_o})\right)$.

We update the elements of $\mathbf{w}_\mathcal{R}$ sequentially. For any two locations ℓ_1 and ℓ_2 in $\mathcal{L}$, if $\ell_1 \in N(\ell_2)$ and is the j-th member of $N(\ell_2)$, then we define b_{ℓ_2,ℓ_1} as the j-th entry of $\mathbf{a}_{N(\ell_2)}$. Let $U(\ell_1) = \{\ell_2 \in \mathcal{R} \,|\, \ell_1 \in N(\ell_2)\}$ and for every $\ell_2 \in U(\ell_1)$, define, $a_{\ell_2,\ell_1} = w(\ell_2) - \sum_{\ell \in N(\ell_2), \ell \neq \ell_1} w(\ell) b_{\ell_2,\ell}$. Then, for $i = 1, 2, \ldots, n$ the full conditional distribution for $\mathbf{w}(\ell_i)$ is $N\left(v(\ell_i)\mu(\ell_i), v(\ell_i)\right)$, where

$$v(\ell_i) = \left(\frac{I(\ell_i)}{\tau^2} + \frac{1}{f_{\ell_i}} + \frac{\sum_{\ell \in U(\ell_i)} b_{\ell,\ell_i}^2}{f_\ell}\right)^{-1} \quad \text{and} \\ \mu(\ell_i) = \frac{y(\ell_i) - \mathbf{x}(\ell_i)'\boldsymbol{\beta}}{\tau^2} I(\ell_i) + \frac{\mathbf{a}'_{N(\ell_i)}\mathbf{w}_{N(\ell_i)}}{f_{\ell_i}} + \sum_{\ell \in U(\ell_i)} \frac{b_{\ell,\ell_i} a_{\ell,\ell_i}}{f_\ell} \,. \tag{3.11}$$

If $U(\ell_i)$ is empty for some ℓ_i, then all instances of $\sum_{\ell \in U(\ell_i)}$ in (3.11) disappear for that $\mathbf{w}(\ell_i)$.

3.5.2 Metropolis step

We update $\boldsymbol{\theta}$ using a random walk Metropolis step. The full-conditional for $\boldsymbol{\theta}$ is proportional to

$$p(\boldsymbol{\theta})p(\mathbf{w}_{\mathcal{R}} \,|\, \boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \times \prod_{i=1}^{n} N\left(\mathbf{w}(\ell_i) \,|\, \mathbf{a}'_{N(\ell_i)}\mathbf{w}_{N(\ell_i)}, f_{\ell_i}\right) \ . \tag{3.12}$$

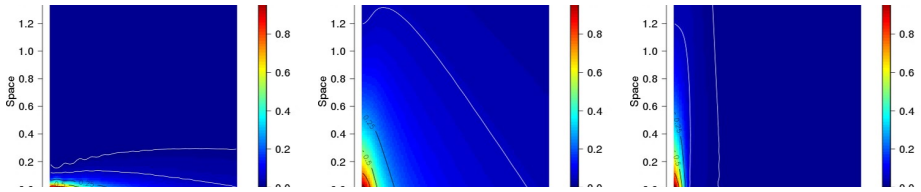
Since none of the above updates involve expensive matrix decompositions, the likelihood can be evaluated very efficiently. The algorithm for updating the parameters of a hierarchical DNNGP model is analogous to the corresponding updates for a purely spatial NNGP model (see Datta et al. (2016)). The only additional computational burden stems from updating the neighbor sets in the adaptive neighbor selection scheme, but even this can be handled efficiently using eligible sets (Algorithm 1). Hence, the number of floating point operations per update is linear in the number of points in $\mathcal{L}$.

3.5.3 Prediction

Once we have computed the posterior samples of the model parameters and the spatio-temporal random effects over $\mathcal{R}$, we can execute, cheaply and efficiently, full posterior predictive inference at unobserved locations and time points. The Gibbs' sampler in Section 3.5.1 generates full posterior distributions of the $\mathbf{w}$'s at all locations in $\mathcal{R}$. Let ℓ_i^* denote a point in $\mathcal{R}$ where the response is unobserved i.e. $I(\ell_i^*) = 0$. We already have posterior distributions of $\mathbf{w}(\ell_i^*)$ and the parameters. We can now generate posterior samples of $\mathbf{y}(\ell_i^*)$ from $N(\mathbf{x}(\ell_i^*)'\boldsymbol{\beta} + \mathbf{w}(\ell_i^*), \tau^2)$. Turning to prediction at a location ℓ outside $\mathcal{R}$, we construct $N(\ell)$ from $E(\ell)$ described in Equation 3.7 for every posterior sample of $\boldsymbol{\theta}$. We generate posterior samples of $\mathbf{w}(\ell)$ from $N(\mathbf{a}'_{N(\ell)}\mathbf{w}_{N(\ell)}, f_\ell)$ and, subsequently, draw posterior samples of $y(\ell)$ from $N(\mathbf{x}(\ell)'\boldsymbol{\beta} + \mathbf{w}(\ell), \tau^2)$.

3.6 Synthetic data analyses

In this section we compare the DNNGP, the full rank GP and low rank Gaussian Predictive Process (Banerjee et al., 2008). We generated observations over a $n = 15 \times 15 \times 15 = 3375$ grid within a unit cube domain. An additional 500 observations used for out-of-sample prediction validation were also located within the domain. All data were



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 3.4: Space-time correlation surface realizations given *true* parameter values in Table 3.1. Correlation contours are provided, with the two outer white lines corresponding to 0.05 and 0.01.

generated using model 6.1 with $\mathbf{x}(\ell)$ comprising an intercept and covariate drawn from $N(0,1)$. The spatial covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ was constructed using an exponential form of the non-separable spatio-temporal covariance function (3.9), viz.,.

$$\frac{\sigma^2}{(a|u|^2+1)^{\kappa}}\exp\left(\frac{-c\|h\|}{(a|u|^2+1)^{\kappa/2}}\right), \tag{3.13}$$

where $u = |t_i - t_j|$ and $h = ||\mathbf{s}_i - \mathbf{s}_j||$ are the time and space Euclidean norms, respectively. By specifying different values of the decay and interaction parameters in $\boldsymbol{\theta} = (\sigma^2, \kappa, a, c)$, we generated three datasets that exhibited different covariance structures. The first column in Table 3.1 provides the three specifications for $\boldsymbol{\theta}$ and Figure 3.4 shows the corresponding space-time correlation surface realizations. As illustrated in Figure 3.4, the three datasets exhibit: 1) short spatial range and long temporal range, 2) long spatial and temporal range, and 3) long spatial range and short temporal range.

For each dataset, model parameters were estimated using: *i*) full Gaussian Process (GP), *ii*) DNNGP with simple neighbor set selection (Simple DNNGP) described in Section 3.4.1, *iii*) DNNGP with adaptive neighbor set selection (Adaptive DNNGP) described in Section 3.4.2, and; *iv*) bias-corrected Gaussian Predictive Process (GPP) detailed in Banerjee et al. (2008) and Finley et al. (2009). DNNGP models were fit using $m = \{16, 25, 36\}$ and the Gaussian Predictive Process model used a regularly spaced grid of $8 \times 8 \times 8 = 512$ knots within the domain.

For all models, the intercept β_0 and and slope regression parameters, β_1, were assigned *flat* prior distributions. The variance parameters were assumed to follow inverse-Gamma prior distributions with $\sigma^2 \sim IG(2,1)$ and $\tau^2 \sim IG(2, 0.1)$. The time and space

decay parameters received uniform priors that were dataset specific: 1) $a \sim U(1, 100)$, $c \sim U(0, 50)$; 2) $a \sim U(300, 700)$, $c \sim U(0, 10)$, and; 3) $a \sim U(1000, 3000)$, $c \sim U(0, 10)$. The prior for the interaction term matched its theoretical support with $\kappa \sim U(0, 1)$.

Candidate model comparison was based on parameter estimates, fit to the observed data, out-of-sample prediction accuracy, and posterior predictive distribution coverage. Goodness-of-fit was assessed using DIC (Spiegelhalter et al., 2002) and posterior predictive loss (Gelfand and Ghosh, 1998). The DIC is reported along with an estimate of model complexity, pD, while the posterior predictive loss is computed as D=G+P, where G is a goodness-of-fit measure and P measures the number of model parameters. Predictive accuracy for the 500 holdout locations was measured using root mean squared prediction error (Yeniay and Goktas, 2002). The percent of holdout locations that fell within the candidate models' posterior predictive distribution's 95% credible interval (CI) was also computed. Inference was based on 15,000 MCMC samples comprising post burn-in samples from three chains of 25,000 iterations (i.e., 5,000 samples from each chain).

Table 3.1 presents parameter estimation and model assessment metrics. With the exception of τ^2 for Dataset 1, the full GP model recovered the parameter values used to generate the datasets, i.e., the 95% CIs cover the *true* parameter values. For the DNNGP models, there was negligible difference among parameter estimates for the 15, 25, and 36 neighbor sets. Hence, we report only the $m = 25$ cases. There was very little difference between the estimates produced by the Adaptive and Simple DNNGP models and, like the full GP model, they captured the *true* mean and process parameters, with the exception of τ^2 for Dataset 1. Given the extremes in the space and time decay in Datasets 1 and 3, we anticipated the Simple DNNGP model—with at most 5 neighbors in any given time point—would not be able to estimate the covariance parameters. Extensive analysis of simulated data, some of which is reported in Table 3.1, suggested the Simple DNNGP model performed as well as the Adaptive DNNGP and full GP models. Goodness-of-fit and out-of-sample prediction validation metrics in Table 3.1 also show the full GP and DNNGP models provided comparable results. In contrast the GPP model did not capture many of the process parameters and provided worse fit and prediction than the GP and DNNGP models. The quality of the GPP results would improve with additional knots. However, computing time would also increase.

Table 3.1: Synthetic data analysis parameter estimates and computing time for the candidate models. Parameter posterior summary 50 (2.5, 97.5) percentiles. Bold indicates estimates with 95% credible intervals that do not include the *true* parameter value.

		GP	GPP knots=512	Adaptive DNNGP m=25	Simple DNNGP m=25
			Dataset 1		
β_0	1	0.99 (0.80, 1.12)	1.02 (0.89, 1.16)	0.97 (0.86, 1.11)	0.97 (0.86, 1.11)
β_1	5	4.99 (4.97, 5.01)	4.98 (4.94, 5.02)	4.99 (4.97, 5.01)	4.99 (4.97, 5.01)
a	50	46.46 (38.02, 67.46)	**16.93 (11.91, 29.17)**	48.37 (37.85, 77.93)	53.18 (35.93, 83.78)
c	25	25.69 (22.00, 29.49)	22.73 (13.53, 34.20)	24.77 (20.55, 28.71)	25.16 (21.91, 29.52)
k	0.75	0.83 (0.61, 0.94)	0.78 (0.39, 0.91)	0.80 (0.57, 0.99)	0.75 (0.53, 0.98)
σ^2	1	**1.13 (1.03, 1.24)**	**0.70 (0.56, 0.92)**	**1.13 (1.03, 1.24)**	**1.14 (1.04, 1.25)**
τ^2	0.1	0.09 (0.07, 0.11)	**0.95 (0.89, 1.02)**	0.09 (0.06, 0.11)	0.09 (0.06, 0.11)
DIC		3700.68	9644.76	3650.55	3567.45
D=G+P		616.90	6444.93	604.72	588.13
RMSPE		0.84	0.95	0.84	0.84
95% CI cover %		95.6	94.6	95.6	96
			Dataset 2		
β_0	1	0.81 (0.48, 1.26)	0.79 (0.26, 1.16)	1.01 (0.57, 1.27)	0.92 (0.58, 1.26)
β_1	5	4.98 (4.96, 5.00)	4.99 (4.97, 5.02)	4.98 (4.96, 5.00)	4.98 (4.96, 5.00)
a	500	352.82 (301.69, 521.64)	583.59 (391.79, 661.36)	480.28 (331.63, 662.12)	410.84 (317.29, 602.21)
c	2.5	2.52 (1.93, 3.13)	**1.67 (1.03, 2.31)**	2.91 (2.49, 3.37)	2.59 (1.91, 3.37)
k	0.5	0.56 (0.44, 0.67)	0.39 (0.26, 0.53)	0.46 (0.36, 0.62)	0.53 (0.42, 0.67)
σ^2	1	1.01 (0.85, 1.31)	1.14 (0.83, 1.77)	0.94 (0.81, 1.10)	1.03 (0.83, 1.32)
τ^2	0.1	0.11 (0.09, 0.13)	**0.44 (0.41, 0.47)**	0.10 (0.08, 0.12)	0.10 (0.08, 0.12)
DIC		3988.36	7091.84	3866.38	3871.09
D=G+P		733.53	2946.94	689.56	687.68
RMSPE		0.53	0.71	0.53	0.53
95% CI cover %		96.4	93	94	93.8
			Dataset 3		
β_0	1	0.94 (0.66, 1.14)	**0.55 (0.32, 0.84)**	0.93 (0.74, 1.17)	0.87 (0.68, 1.12)
β_1	5	4.98 (4.96, 5.00)	4.98 (4.95, 5.02)	4.98 (4.96, 5.00)	4.98 (4.96, 5.00)
a	2000	1214.02 (1008.23, 2141.16)	1590.77 (1151.78, 2118.63)	1635.46 (1046.76, 2922.59)	1495.94 (1019.16, 2751.17)
c	2.5	2.38 (1.79, 2.95)	**1.36 (0.73, 2.16)**	2.25 (1.62, 2.81)	2.27 (1.60, 2.97)
k	0.95	0.91 (0.72, 0.98)	**0.68 (0.40, 0.90)**	0.71 (0.46, 0.98)	0.81 (0.58, 0.97)
σ^2	1	1.03 (0.86, 1.35)	0.91 (0.67, 1.83)	1.09 (0.89, 1.44)	1.07 (0.87, 1.42)
τ^2	0.1	0.11 (0.09, 0.13)	**0.68 (0.62, 0.74)**	0.11 (0.09, 0.14)	0.11 (0.09, 0.14)
DIC		4210.71	8463.33	4214.68	4223.88
D=G+P		765.89	4562.21	769.13	768.59
RMSPE		0.78	0.92	0.77	0.77
95% CI cover %		92.8	91.4	95.6	95.6
CPU (min)		7646.96	856.54	496.12	430.88

The last row in Table 3.1 provides the CPU time required for each candidate model to generate 25,000 MCMC samples for the $n = 3375$ observations. Even with the substantial dimension reduction, the GPP model required about twice the CPU time as the DNNGP models. Compared to the full GP model, the DNNGP models provided substantial computational advantages while delivering comparable results.

3.7 Analysis of Airbase and LOTOS-EUROS CTM data

We consider the model in Equation 4.3, where $y(\ell_i)$ is a square-root transformed measurement of PM_{10} at space-time coordinate ℓ_i, $x(\ell_i)$ is the coinciding square-root transformed output from the LOTOS-EUROS CTM. Given the large dimension of the dataset, $n = N \times M = 308 \times 730 = 224{,}840$, the spatio-temporal random effects were modeled as a DNNGP prior derived from a zero-centered GP with the non-separable spatio-temporal covariance function (3.13). Exploratory analysis—consisting of semivariogram plots and autocorrelation function plots for simple ordinary least square model residuals—helped guide choice of prior and hyper-parameters for the variance and decay parameters. Specifically, $\sigma^2 \sim IG(2, 1)$, $\tau^2 \sim IG(2, 0.1)$, $a \sim U(0.1, 5)$, and $c \sim U(0.01, 0.5)$, with κ fixed at 0.5.

Candidate models included the *i*) LOTOS-EUROS CTM, *ii*) simple linear regression model with no spatio-temporal effects, i.e., $w(\ell) = 0$, and *iii*) Adaptive and Simple DNNGP with $m = \{16, 25, 36\}$. Following Section 3.6, candidate model goodness-of-fit to the observed data was assessed using DIC and GPD, whereas predictive performance was assessed using RMSPE and 95% posterior predictive CI coverage rate for out-of-sample prediction. The holdout set comprised blocks of five days per station—five days of continuous observations were withheld at random from each station's 730 day time series.

Additionally, prediction using the Adaptive and Simple DNNGP models for a 25% holdout set selected from April 1-14, 2009 was compared with results from Hamm et al. (2015) who considered time invariant spatial regression models for the same two-week period and comparable prediction validation approach.

A subset of analysis results are given in Table 3.2. Parameter estimates for the model intercept and regression slope coefficient associated with the CTM output are consistent

Table 3.2: PM_{10} analysis parameter posterior 50 (2.5, 97.5) percentiles, model fit and prediction metrics, and run time for 25,000 MCMC samples.

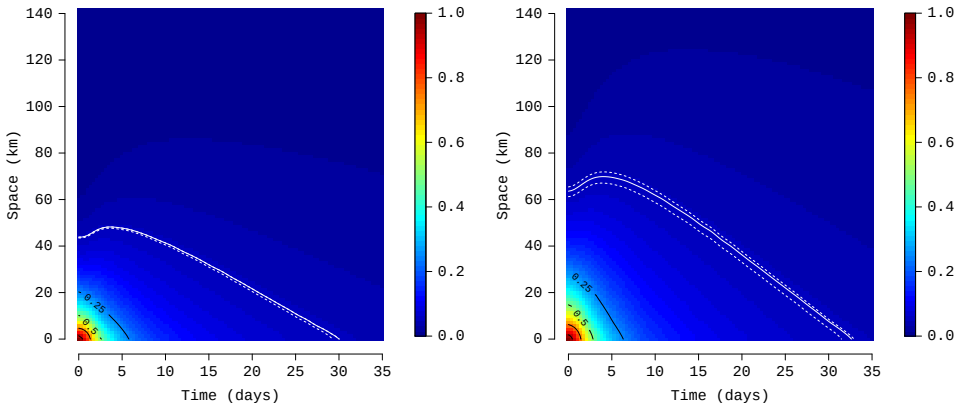
		Adaptive			Simple
Parameter	Non space-time	m=16	m=25	m=36	m=36
β_0	1.66 (1.64, 1.68)	2.56 (2.53, 2.59)	2.62 (2.59, 2.65)	2.61 (2.58, 2.64)	2.64 (2.61, 2.68)
β_1	0.76 (0.75, 0.76)	0.47 (0.46, 0.47)	0.45 (0.44, 0.46)	0.45 (0.44, 0.46)	0.44 (0.43, 0.45)
a	–	0.57 (0.57, 0.57)	0.44 (0.44, 0.44)	0.46 (0.46, 0.46)	0.37 (0.37, 0.39)
c	–	0.08 (0.08, 0.08)	0.07 (0.07, 0.07)	0.07 (0.07, 0.07)	0.05 (0.05, 0.05)
σ^2	–	1.49 (1.48, 1.51)	1.64 (1.62, 1.66)	1.56 (1.54, 1.58)	2.06 (2.01, 2.11)
τ^2	1.48 (1.47, 1.48)	0.12 (0.12, 0.12)	0.14 (0.14, 0.14)	0.14 (0.14, 0.14)	0.15 (0.15, 0.16)
P_D	2.75	110266.2	122466.2	111190.6	103038.3
DIC	586,135.8	279077.3	265720.6	277383.9	286,922.9
G	432811.9	11538.63	8707.79	11249.11	13521.63
P	268036.7	40994.19	36711.28	40532.25	43728.23
D	700848.6	52532.82	45419.07	51781.37	57249.86
RMSPE	12.75	8.28	8.24	8.2	8.11
95% CI cover %	93.4	93.33	93.06	93.15	92.86
CPU (min)	–	6182.89	15681.8	27660.5	25819

across the candidate models. For an accurate CTM it would be expected that $\beta_0 \approx 0$ and $\beta_1 \approx 1$. The finding that $\beta_0 > 0$ and $0 < \beta_1 < 1$ corroborate previous findings that showed the CTM consistently underestimates PM_{10} (Stern et al., 2008; Hamm et al., 2015). The spatial and temporal decay parameters differed between the Adaptive and Simple DNNGP models. Figure 3.5 provides correlation surfaces generated using posterior median values of a and c from the $m = 36$ Adaptive and Simple DNNGP models (using values given in Table 3.2). The 0.05 correlation contour on these surfaces suggest the Simple model estimates a moderately longer spatial and temporal range, i.e., ∼60 km and ∼33 days, versus ∼45 km and ∼30 days for the Adaptive model. Within a given DNNGP neighbor selection algorithm there is only marginal difference between the covariance parameters estimates when comparing m of 25 and 36. Neighbor sets of less than 25 provided consistently larger temporal decay parameter estimates, i.e., shorter temporal correlation estimates, although even with such few neighbors the models seemed to produce consistent estimates of the spatial decay.

The spatial range of 45 to 60 km is an order of magnitude less than that observed by Hamm et al. (2015), who estimated median spatial ranges of 500 to 1500 km. This is attributed to the inclusion of temporal correlation in the model, which itself accounts for a large amount of the residual spatial structure. The temporal range is physically reasonable considering the life-time of PM_{10} is in the order of days and its variability is driven by alternating synoptic meteorological conditions, with certain conditions usually lasting for several days to weeks.

Across all candidate models the Adaptive with m=25 provided the lowest values of DIC and D suggesting improved fit to the observed data. This improved fit did not correspond to increased out-of-sample prediction accuracy. Rather, RMSPE consistently decreased with increasing number of neighbors within the Adaptive and Simple model sets. The smallest RMSPE was achieved using the simple neighbor selection with m=36. All models achieved reasonable coverage rates.

Figure 3.6 illustrates the observed and candidate model fitted/predicted PM_{10} for three stations. These figures are representative of other stations and show *i*) the downward bias in CTM output, *ii*) improved fit and prediction with the addition of spatio-temporal random effects over non-spatial regression, and *iii*) appropriate widening of CIs for missing station observations.

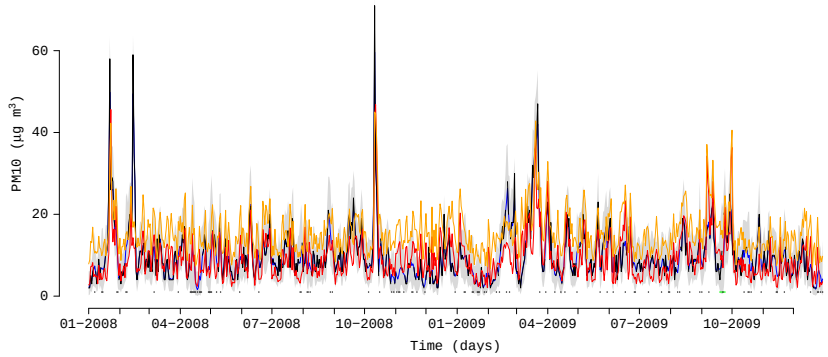


(a) Adaptive m=36 (b) Simple m=36

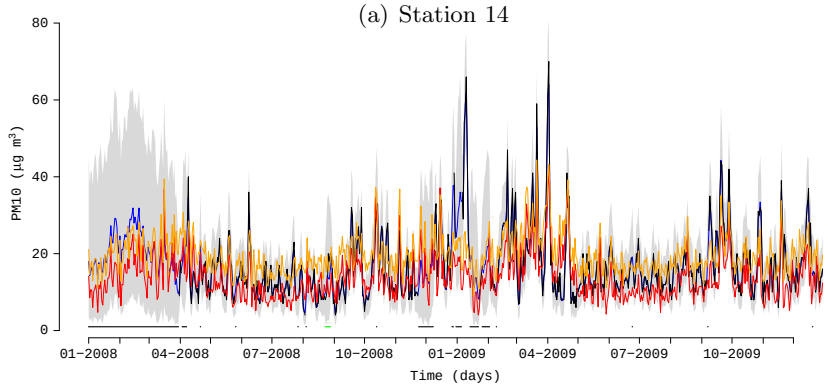
Figure 3.5: Space-time correlation posterior distribution median surfaces. Median (white lines) and associated 95% credible intervals (dotted white lines) for correlation contours of 0.05.

Table 3.3 provides out-of-sample prediction validation metrics for the non space-time and DNNGP Adaptive and Simple models that can be compared with April 1-14, 2009 holdout validation metrics presented in Hamm et al. (2015, Table 1). Compared to the time invariant (day specific) space-varying intercept (SVI) and space-varying coefficients (SVC) models considered in Hamm et al. (2015), the DNNGP models' RMSPE and bias are lower (more accurate, less biased) while the R^2 values are comparable. We also added results for the simple linear regression (SLR) model in the first column of Table 3.3. The simple linear regression model does not consider spatio-temporal effects nor does it consider a time varying intercept (unlike the day specific results presented in Hamm et al. (2015)) which may explain the poor predictive performance—it is more meaningful to compare the DNNGP model prediction metrics to the days specific metrics presented in Hamm et al. (2015).

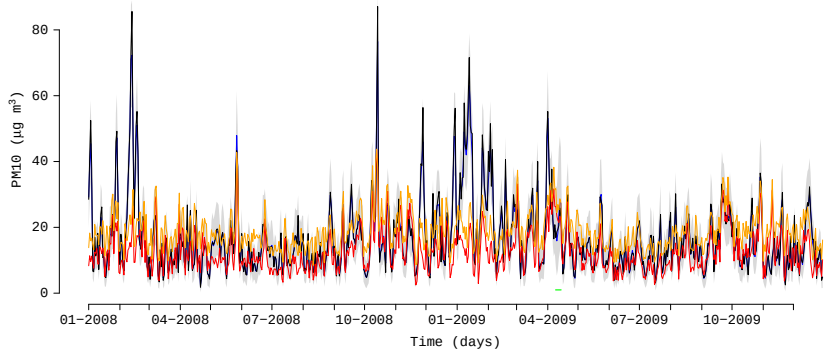
In addition to these prediction metrics, maps of posterior predictive summaries at CTM output locations are key inputs to pollution monitoring and mitigation programs. For example, Figure 3.7 provides maps of the posterior predictive prediction median and the probability of exceeding the 50 μg m^{-3} regulatory threshold for two example dates. These dates were also examined in Hamm et al. (2015, Figure 8) and the resulting maps



(a) Station 14



(b) Station 47

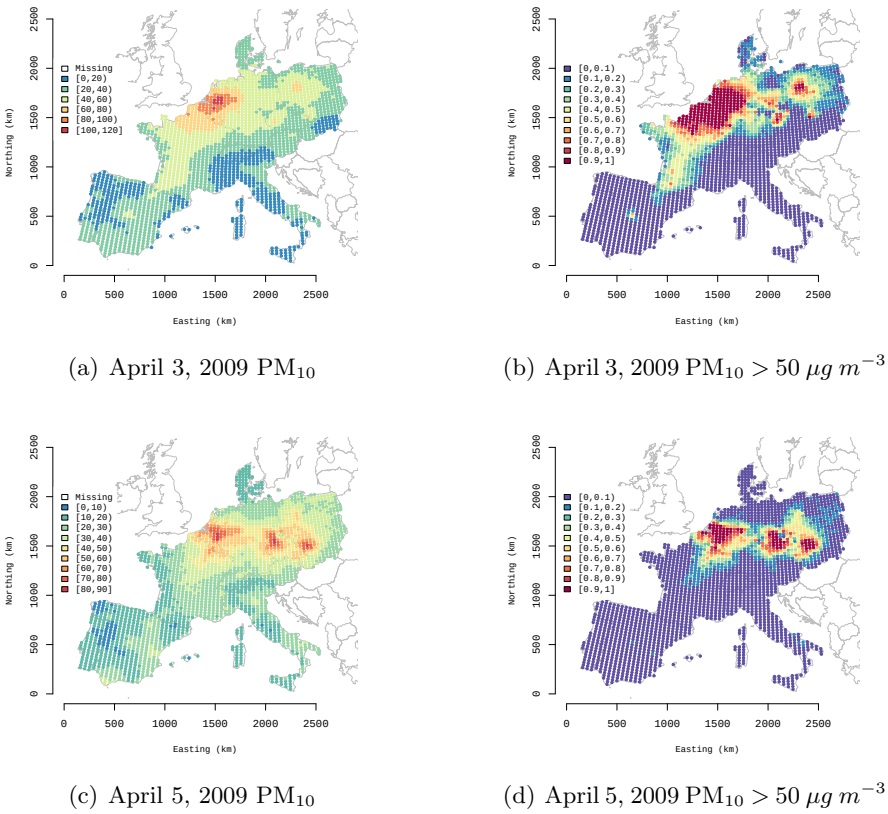


(c) Station 151

Figure 3.6: Fitted and observed PM_{10} for several example stations. Lines correspond to PM_{10} observed (black), CTM output (red), non space-time, regression (orange), and $m = 36$ Adaptive DNNGP (blue) with associated 95% CI band (gray). Prediction assessment holdout and actual missing observations are indicated with green and black points respectively.

Table 3.3: April 1-14, 2009 25% holdout set prediction summary for comparison with time invariant spatial regression models presented in (Hamm et al., 2015, Table 1).

		Adaptive		Simple	
	SLR	m=25	m=36	m=25	m=36
RMSPE	8.48	4.97	5.05	5.06	5.04
Bias	0.71	0.20	0.20	0.23	0.22
R^2	0.14	0.69	0.68	0.68	0.68



(a) April 3, 2009 PM_{10}

(b) April 3, 2009 $PM_{10} > 50\ \mu g\ m^{-3}$

(c) April 5, 2009 PM_{10}

(d) April 5, 2009 $PM_{10} > 50\ \mu g\ m^{-3}$

Figure 3.7: Predicted PM_{10} and probability of exceeding 50 $\mu g\ m^{-3}$ for two example dates.

are directly comparable. The DNNGP, Figure 3.7, and SVC maps in Hamm et al. (2015) show broadly similar patterns, although there are some differences. For example the high pollution over western France and northern Spain on April 3, 2009 is captured more

clearly by Hamm et al. (2015). The SVI and SVC models in Hamm et al. (2015) did not account for temporal correlation over days—clearly not an accurate assumption. In contrast the DNNGP models smooth over days, which can provide improved predictive performance although the details of highly dynamic events may be less well captured than by the daily specific models used in Hamm et al. (2015).

The last row in Table 3.2 provides the CPU time for delivering 25,000 MCMC iterations. As detailed in Section 3.4.2 particular components of the algorithm are easily distributed across multiple CPUs. In particular, partitioning the update of $w(\ell_i)$'s across multiple CPUs yields substantial computational gains. The DNNGP samplers were implemented in C++ and leveraged OpenMP (Dagum and Menon, 1998) and Intel Math Kernel Library's (MKL) threaded BLAS and LAPACK routines for matrix (Intel, 2015). Running on a single CPU the Adaptive m=25 model would require approximately 260 hours. However, when distributed across a 10-core Xeon CPU the total run time was approximately 24 hours.

3.8 Conclusion

We have addressed the problem of modeling large spatio-temporal datasets, specifically for settings where full inference (with proper accounting for uncertainty) is required at arbitrary resolutions. We presented a new class of dynamic nearest-neighbor Gaussian Process (DNNGP) models over a continuous space-time domain. The DNNGP is a legitimate Gaussian process whose realizations over finite sets enjoy sparse precision matrices, thereby accruing massive computational savings in terms of storage and flops. The DNNGP depends upon the conditional independence of the random effects given its neighbors. We used the strength of a correlation function to construct a parametric distance metric in a spatio-temporal domain. Using monotonicity of covariance functions we showed that it is possible to update neighbor sets using a scalable search algorithm and outlined the steps of a Gibbs' sampler that avoids expensive matrix decompositions and is linear in the number of measurements in terms of storage and flops.

Analyses combining European CTM outputs and observed data has, to date, focused mainly on spatial analysis per day (Denby et al., 2008, 2010; Hamm et al., 2015), few studies implement full space-time geostatistical models, e.g., Gräler et al. (2011), and

none consider such a long time series. The work presented in this paper focuses on DNNGP development to facilitate novel analyses of spatially-indexed time-series data such as PM_{10} concentrations. Here, in addition to improved predictive performance, inference on model covariance parameters provided insight into space-time structures not captured by the LOTOS-EUROS CTM. Whilst previous analyses of individual days had shown strong residual spatial structure, analysis of this long time-series with explicit time correlation parameters reveals the residual temporal structure dominates. The temporal range is physically reasonable considering the life-time of PM_{10} is in the order of days and its variability is driven by alternating synoptic meteorological conditions with certain conditions usually lasting for several days to weeks.

Reproducing the observed variability with a CTM remains challenging, especially for episodic conditions which are associated with particular (stagnant) meteorological conditions or occasional large emissions from, e.g., large wild fires (R'Honi et al., 2013) or dust events (Birmili et al., 2008). A particular issue to be resolved is the lack of detail in the anthropogenic emission variability. This variability is prescribed using static emission profiles for the month of the year, day of the week, and hour of the day. Further detailing through inclusion of meteorological effects may improve the modeling (Mues et al., 2014) and remove the monthly signature found in this analysis.

The type of analysis that is performed depends on the study objective. Analysis of individual days is important for the study of individual air pollution events and the associated performance of the CTM (Hamm et al., 2015). The analysis presented in this paper affords a different perspective by identifying long-term space-time structures that offer insight into the performance of the CTM. The DNNGP also yields more accurate predictions than previous studies of these same data.

Apart from massive scalability, the DNNGP retains the versatility of process-based modeling and can be used as a sparsity-inducing proper prior in any Bayesian hierarchical model designed to deliver full inference at arbitrary spatio-temporal resolutions for massive spatio-temporal datasets. We have developed DNNGP assuming an isotropic non-stationary spatio-temporal covariance structure. However, it can also be potentially extended to certain classes of non-stationary space-time covariances. Even more generally, the DNNGP can be used for any spatio-temporal random effect in the second stage of specification in hierarchical models for non-Gaussian responses. Full posterior

distributions for the underlying spatio-temporal process are available at any arbitrary location and time point. Thus, DNNGP can potentially be deployed for statistical downscaling of spatio-temporal datasets obtained at coarser resolutions (e.g. climate downscaling).

Chapter 4

Directed Acyclic Graph Autoregressive Models for Areal Datasets

4.1 Introduction

Epidemiological data for different disease rates are often recorded as aggregated disease counts over entire geographical regions like states, counties or other administrative units. Accurate identification of trends and factors associated with the disease requires accounting for the spatial dependence among the regions. However, unlike the situation of chapters 2 or 3, inference for these areal datasets is sought at a coarser resolution, over entire geographical regions instead of isolated locations.

If the dataset contains several replicate observations at each region, then assuming that there are k regions, one can view each replicate as a k-dimensional multivariate vector. Subsequently the sample covariance matrix calculated using the replicates will provide an estimate of the dependence among the different regions. If k is large, the problem reduces to high-dimensional covariance or precision matrix estimation. There exists a vast inventory of statistical methods for estimating such high-dimensional graphical models. Common examples include banding (Wu and Pourahmadi, 2003; Bickel and Levina, 2008b), tapering (Cai et al., 2010), thresholding (Bickel and Levina, 2008a;

El Karoui, 2008; Rothman et al., 2009) and penalization (Meinshausen and Buhlmann, 2006; Friedman et al., 2007; Xue et al., 2012) among others. These methods can be used to infer about the marginal or conditional dependencies among the different regions without using any geographical information about the regions.

However, most areal datasets contain a single count per region. In absence of replicates, none of the aforementioned methods can be used and one needs to use the information offered by the geography of the regions to model the spatial dependence. There are two ways of modeling spatial dependence for such areal datasets. The first approach represents each geographical region with a single location (typically the centroid of the region). This reduces the problem of capturing dependence over discrete regions to the well-studied problem of modeling spatial dependence in a continuous domain. Consequently, Gaussian Process based approaches become relevant once again. However, this approach has two problems. Firstly, for large areal data comprising several regions, one runs into the computational issues that accompany GP based modeling. Chapter 2 demonstrates that this can be alleviated using a NNGP derived from the original GP. The second problem is that the choice of assigning a single location to an entire region is ad hoc and different choices of the representative locations like geographical centroid or population weighted centroid, may lead to different results.

The second approach visualizes the geographical domain as an undirected graph with a vertex at each region and an edge between two vertices if the corresponding regions share a geographical border. This creates well-defined neighbors for each region, which are used to define the joint or conditional distribution. For example, the widely used conditional autoregressive (CAR) model (Besag, 1974; Clayton and Bernardinelli, 1992) incorporates the underlying neighborhood structure in specifying the full conditional distribution for each observation. Throughout the text we adopt the convention that $N(\alpha, \Delta)$ denotes normal distribution with mean α and precision Δ, both in a univariate or multivariate context. If w_1,w_2,...,w_k denotes the observations (usually random effects) at k regions (vertices) and $i \sim j$ implies regions i and j are neighbors, then a CAR model specifies the full conditional distributions as follows:

$$w_i \,|\, w_{-i} \sim N(\frac{1}{n_i} \sum_{j \,|\, i \sim j} w_j, \tau_w n_i) \tag{4.1}$$

where w_{-i} denotes the vector of observations leaving out the i^{th} one and n_i denotes the

number of neighbors of the i^{th} location.

The full conditional means for the CAR models depend on the adjacency matrix of the underlying graph and are averages of neighboring observations. This construction, although intuitive, yields an improper joint distribution of the w_i's and is hence referred to as the intrinsic or improper CAR (ICAR) model. From (4.1) the joint distribution of $w = (w_1, w_2, \ldots, w_k)'$ can be derived as

$$w \sim N(0, \tau_w(D - A)) \tag{4.2}$$

where $A = ((a_{ij}))$ is the adjacency matrix of the neighborhood graph, i.e., $a_{ij} = 1 \iff i \sim j$, and $D = diag(n_1, n_2, \ldots, n_k)$. It is easy to see that $Q_{CAR} = D - A$ is singular as the sum of its entries in each row or column adds up to zero. Although this renders the CAR ineligible to be a model for the response in a conventional setup, the CAR distribution can still be used as a prior for spatial random effects in the second stage of specification. To elucidate, if y_i denote the observed response at region i and x_i denotes the corresponding covariate vector, then a hierarchical model can be specified as

$$\prod_{i=1}^{k} N(y_i \,|\, x_i'\beta + w_i, \tau_e)\; N(w \,|\, \tau_w Q_{CAR})\; p(\beta, \tau_w, \tau_e) \tag{4.3}$$

where $p(\beta, \tau_w, \tau_e)$ is some prior for the parameters. Note that despite the impropriety of CAR, the posterior $w \,|\, y = (y_1, y_2, \ldots, y_k)'$ remains proper. Also Hodges (2013) showed how the CAR can be used in a mixed linear model setup using the eigenvalue decomposition of the precision matrix. So, although the impropriety initially seems uncomfortable, there is a usually a way around it.

The eigen-vector corresponding to the zero eigen-value of a CAR precision matrix is the vector of ones. This implies that the CAR tends to smooth the random effects toward a constant. This feature of a CAR model is highly undesirable as two spatial effects may correspond to two very different locations with different neighborhood structures and should be shrunk differentially. Both this and the impropriety of the CAR model are usually fixed by modifying the full conditional mean to $E(w_i \,|\, w_{-i}) = \frac{\rho}{n_i} \sum_{j \,|\, i \sim j} w_j$. This yields a joint distribution $w \sim N(0, \tau_w(D - \rho A))$ which is proper for a certain range of ρ. However the proper CAR also has known problems. First, the full conditional mean of the proper CAR is not intuitively as meaningful as that of a CAR model. Secondly,

it is difficult to interpret the parameter ρ as it has been observed that even very high values of ρ induces only modest spatial correlation among the observations (see Banerjee et al., 2014, for a discussion on this). Furthermore, Wall (2004) used US state level SAT scores to show that even negative values of ρ may lead to positive correlation among neighboring states. Assuncao and Krainski (2009) used matrix algebra and Markov chain results to reveal that these oddities are a general feature of proper and improper CAR models arising out of the precision specification.

The second popular model, based on the graph adjacency matrix, is the Simultaneous Autoregressive (SAR) model (Whittle, 1954). Instead of taking the conditional route, the SAR model proceeds by simultaneously modeling the random effects as:

$$w_i = \sum_{i \sim j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k \tag{4.4}$$

where $\epsilon_i \overset{\text{ind}}{\sim} N(0, \tau_i)$ are errors independent of w. Defining $B = ((b_{ij}))$ and $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$, the set of equations in (4.4) implies that $w = Bw + \epsilon$ where $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_k)' \sim N(0, F)$. If $I - B$ is non-singular, then this effectuates the following joint distribution

$$w \sim N(0, (I - B)' F (I - B)). \tag{4.5}$$

A common choice is to define $b_{ij} = I(i \sim j)/n_i$ where $I(\cdot)$ denotes the indicator function. However, unlike the CAR model, this does not ensure that the conditional expectation $E(w_i \,|\, w_{-i})$ is simply the average of its neighbors. In fact, the full conditional distributions for the SAR model lack any convenient formulation. For large areal datasets, this leads to computational roadblocks. Like the CAR model, the SAR model as above is also improper and there is a proper SAR model with b_{ij} defined as $\rho I(i \sim j)/n_i$. For $\rho \in (-1, 1)$, $I - B$ is non-singular thereby curing the impropriety. However, like the proper CAR, the proper SAR faces similar issues regarding interpretation of the parameter ρ. Many of the issues of the proper CAR discussed in Wall (2004) are also inherited by the proper SAR model.

Beyond the CAR and SAR models, the inventory of covariance models for areal datasets is very limited. Leroux et al. (2000) and MacNab and Dean (2000) extended the CAR model by accommodating over-dispersion alongside spatial information. They proposed using the precision matrix $\lambda(D - A) + (1 - \lambda)I$ where $\lambda \in [0, 1]$ controls the

degree of dependence among the regions. Observe that, for a regular graph where all vertices have same number of neighbors d, $D = dI$. In that case, $\lambda(D - A) + (1 - \lambda)I$ can be rewritten as $\frac{1+(d-1)\lambda}{d}(D - \rho^* A)$ where $\rho^* = \frac{d\lambda}{1+(d-1)\lambda}$. Thus if the the numbers of neighbors of the vertices do not vary greatly, this approach is somewhat similar to the proper CAR and is encumbered by the aforementioned issues of proper CAR models.

In this chapter, we propose a new way of constructing precision matrices for areal data. Instead of modeling the precision matrix directly, we model the precision matrix using a sparse Cholesky decomposition. The sparse weights for the Cholesky factors are selected in a similar fashion as the CAR or SAR model. However, unlike CAR or SAR, the proposed covariance matrix is guaranteed to be positive definite. It can therefore be used directly to model the response instead of modeling latent random effects.

The CAR or the SAR model uses the adjacency matrix of the undirected graph created from the areal units. In contrast, modeling the Cholesky factor essentially uses a directed acyclic graph created from the original undirected graph. As directed graphs depend on the ordering of the vertices (regions), so does the Cholesky factor. This is undesirable as spatial regions do not have any natural ordering. We remedy this situation by averaging over all possible directed acyclic graphs that can be created from an undirected graph. The resulting precision matrix is order free and available in closed form. We refer to this model as the Directed Acyclic Graph Autoregressive (DAGAR) model. Unlike CAR and SAR, DAGAR precision matrices are positive definite. Furthermore, DAGAR does not require any additional parameters that are difficult to interpret and is only dependent on the adjacency matrix. DAGAR precision matrices are also sparse and, therefore, is suitable to model very large areal datasets. We also develop a computationally efficient Gibbs sampler for hierarchical DAGAR models. We compare the performance of CAR and DAGAR models using simulated datasets on regular and irregular graphs.

The rest of the chapter is organized as follows. Section 4.2 elaborates on the construction of precision matrices based on directed acyclic graphs. Section 4.3 extends this to an order-free Directed Acyclic Graph Autoregressive (DAGAR) model and discusses theoretical properties of DAGAR models. Section 4.4 provides an efficient Gibbs sampler for hierarchical DAGAR models. Simulation experiments are reported in Section 4.5. Finally, Section 4.6 ends the chapter with a brief overview of the work done and

pointers to future research.

4.2 Modeling Cholesky Factors

We assume without loss of generality that all the areal locations are connected (i.e., one island). Models for non-connected datasets with multiple islands will be a simple extension. We start with a connected graph $\mathcal{G} = (V, E)$ with the locations as vertices V and edges E between neighbors. For notational convenience we denote the i^{th} location simply by i. Let $A = ((a_{ij}))$ denote the adjacency matrix for this undirected graph. We begin by specifying simultaneous distributions similar to the SAR model in (4.4) i.e., $w_i = \sum_{i \sim j} b_{ij} w_j + \epsilon_i$. However, we additionally impose the assumption that $b_{ij} = 0$ if $j > i$. The simultaneous equations can now be written as:

$$\begin{aligned}
w_1 &= \epsilon_1 \\
w_2 &= I(2 \sim 1) b_{21} w_1 + \epsilon_2 \\
w_3 &= I(3 \sim 1) b_{31} w_1 + I(3 \sim 2) b_{32} w_2 + \epsilon_3 \\
&\vdots \\
w_k &= I(k \sim 1) b_{k1} w_1 + I(k \sim 2) b_{k1} w_2 + \ldots + I(k \sim k-1) b_{k,k-1} w_{k-1} + \epsilon_k
\end{aligned} \tag{4.6}$$

and $B = ((b_{ij}))$ becomes lower triangular matrix. If $F = diag(\tau_1, \tau_2, \ldots, \tau_k) = Cov(\epsilon)^{-1}$, then $w \sim N(0, V'FV)$. Here $V = I - B$ is lower triangular with one on the diagonals, which guarantees that $V'FV$ is positive definite.

When the dataset contains replicates, there is substantial literature on modeling sparse Cholesky factors for high-dimensional precision matrices (Wu and Pourahmadi, 2003; Huang et al., 2006; Rothman et al., 2008; Levina et al., 2008; Wagaman and Levina, 2009; Lam and Fan, 2009). These approaches do not need any prior information about the sparsity of the directed acyclic graph and learn it from the data. On the other hand, if a spatial dataset is observed over a large number of locations and time points, the Nearest Neighbor Gaussian Processes developed in Chapters 2 and 3 demonstrate how to construct sparse Cholesky factors that can reconstruct the original GP covariance with very high accuracy. However, NNGPs are derived from an original GP covariance.

Most areal datasets lack replication that would permit use of the high-dimensional learning methods. Furthermore, there is no well defined covariance matrix analogous to

GP covariances for areal observations from which one can derive a sparse approximation (recall that both CAR and SAR are improper distributions). Consequently, we cannot estimate B and F in their most general form and need to make simplifying assumptions similar to the CAR and SAR model. We assume that $E(w_i \,|\, w_1, w_2, \ldots, w_{i-1})$ has equal weights for each w_j such that $j \sim i$ and $j < i$. To be specific, we assume $b_{ij} = 1/m_i$ so that $w_i = \frac{1}{m_i} \sum_{j=1}^{i-1} I(i \sim j) w_j$. We also allow heteroscedasticity among the ϵ_i's i.e $\epsilon \sim N(0, \tau_i)$. At this point, we do not make any further assumptions about m_i and τ_i except that they are positive numbers.

Unlike covariance or precision matrices which remain invariant under different orderings of the multivariate vector, Cholesky factors depend on the ordering of the observations. The construction in (4.6) assumes a specific ordering which we now generalize to any other ordering. Let $\pi = \{\pi(1), \pi(2), \ldots, \pi(n)\}$ denote any predetermined ordering of the data locations and π^{-1} denote its corresponding inverse permutation. Under this ordering, for any $i \neq \pi(1)$, we define its 'past' observations $w_{<i,\pi}$ as the collection $\{w_j \,|\, \pi^{-1}(j) < \pi^{-1}(i)\}$ and its set of directed neighbors $N_\pi(i)$ as the set of neighbors of i in $\mathcal{G}$ which belong to its 'ordered past' i.e.

$$N_\pi(i) = \{j \,|\, i \sim j \text{ and } \pi^{-1}(j) < \pi^{-1}(i)\} \tag{4.7}$$

Note that directed neighbors are not commutative. In fact, $j \in N_\pi(i)$ implies $i \notin N_\pi(j)$ and vice versa. Let E_π denote the collection of directed edges from all members of $N_\pi(i)$ to i for every $i \neq \pi(1)$. We now have a directed acyclic graph $\mathcal{D}_\pi = (V, E_\pi)$. Figure 4.1 shows the construction of directed graphs for two different permutations.

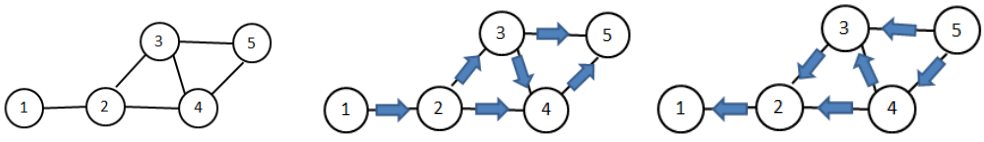


Figure 4.1: Undirected graph (left), D_π with $\pi(1,2,3,4,5) = 1,2,3,4,5$ (middle) and D_π with $\pi(1,2,3,4,5) = 5,4,3,2,1$ (right)

We now specify our model on a single directed graph D_π as follows:

$$\begin{aligned} w_{\pi(1)} &\sim N(0, \tau^{(\pi)}_{\pi(1)}) \\ w_i \,|\, w_{<i,\pi} &\sim N(\tfrac{1}{m_i^{(\pi)}} \textstyle\sum_{j \in N_\pi(i)} w_j, \tau_i^{(\pi)}),\ i \neq \pi(1) \end{aligned} \tag{4.8}$$

The model specification is very similar to the SAR model in (4.4) except that the conditioning sets now depend only on the 'past' instead of on all other locations. As a result, the conditional mean is a weighted sum of the directed neighbors instead of all neighbors. Thus, this is a spatial analogue of an autoregressive model in a time series context where time creates a natural ordering and observation at every time point is regressed on its directed neighbors, i.e., observations at past timepoints. In the CAR model, the joint distribution is constructed from the full conditionals by application of Brooke's lemma, which leads to the impropriety. The advantage of conditioning sets based on a directed acyclic graph is that the joint density is simply a product of the conditional densities. Also, as shown earlier, the precision matrix is guaranteed to be positive definite. We rephrase the result for a general permutation in Theorem 1.

Theorem 1. *Let P_π denote the permutation matrix corresponding to the permutation π i.e. $P_\pi(x_1, x_2, \ldots, x_k)' = (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(k)})'$ for any k-dimensional vector x. Then, $P_\pi w \sim N(0, Q_\pi)$ where $Q_\pi = V_\pi' F_\pi V_\pi$ where $F = diag(\tau^\pi_{\pi(1)}, \tau^\pi_{\pi(2)}, \ldots, \tau^\pi_{\pi(k)})$ and V_π is a lower triangular matrix such that*

$$(V_\pi)_{ij} = \begin{cases} 1 & \text{if } i = j \\ -\frac{1}{m^{(\pi)}_{\pi(i)}} & \text{if } j < i \text{ and } \pi(j) \sim \pi(i) \\ 0 & \text{otherwise} \end{cases}$$

It is clear from Theorem 1 that Q_π is positive definite with $\det(Q_\pi) = \prod_{i=1}^k \tau_i^\pi$. So, w has the proper Gaussian distribution given by

$$w \sim N(0, P_\pi' Q_\pi P_\pi) \tag{4.9}$$

Hence, switching from the undirected graph $\mathcal{G}$ to the directed graph D_π leads to a proper Gaussian distribution. However, the precision matrices and hence the distributions in (4.9) depends on the ordering of the locations. This is highlighted by using the superscript π in m_i and τ_i in Theorem 1. Areal datasets usually do not come with

any natural ordering and there are $k!$ possible orderings. Hence, picking any particular ordering over another will be arbitrary. Furthermore, as $k!$ becomes extremely large, even for very small k, Bayesian model averaging type approaches will fail to explore even a small fraction of all possible orderings.

4.3 Order-free model

Let Q denote the average of the precision matrices in (4.9) over all permutations π, i.e.,

$$Q = \frac{1}{k!} \sum_{\pi} P_\pi' Q_\pi P_\pi \tag{4.10}$$

The matrix Q is free of any ordering. Additionally, since it is the average of positive definite matrices, it is also positive definite. However, Q has $(2k)k!$ unknown parameters $\{m_i^{(\pi)}, \tau_i^{(\pi)} \,|\, i = 1, 2, ..., k, \pi \in \Pi_k\}$ where Π_k is the set of all permutations of $\{1, 2, \ldots, k\}$. To derive a closed form expression for Q we further assume that $m_i^{(\pi)} = m_i$ and $\tau_i^{(\pi)} = \tau_i$ for all $\pi \in \Pi_k$. The following theorem gives the closed form expression for Q.

Theorem 2. *Let $r \sim (i, j)$ imply that r is a common neigbhor of i and j and $i \approx j$ imply that i and j share at least one common neighbor, i.e., they are second order neighbors. Then*

(a) $Q = \Lambda + R$ where

$$\Lambda_{ij} = \begin{cases} \tau_i \text{ if } i = j \\ -\frac{I_{i\sim j}}{2}\left(\frac{\tau_i}{m_i} + \frac{\tau_j}{m_j}\right) \text{ if } i \neq j \end{cases} \quad \text{and } R_{ij} = \begin{cases} \frac{1}{2}\sum_{r\sim i} \frac{\tau_r}{m_r^2} \text{ if } i = j \\ I(i \approx j)\left(\frac{1}{3}\sum_{r\sim(i,j)} \frac{\tau_r}{m_r^2}\right) \text{ if } i \neq j \end{cases}$$

(b) If m denotes the maximum degree of the graph $\mathcal{G}$, then Q has at most $k(1+m+m^2)$ non-zero elements.

Proof. From Theorem 1 we see that

$$(P_\pi' Q_\pi P_\pi)_{ii} = \tau_i + \sum_{r:i\in N_\pi(r)} \frac{\tau_r}{m_r^2}$$

$$(P_\pi' Q_\pi P_\pi)_{ij} = -I(j \in N_\pi(i))\frac{\tau_i}{m_i} - I(i \in N_\pi(j))\frac{\tau_j}{m_j} + \sum_{r:\{i,j\}\subset N_\pi(r)} \frac{\tau_r}{m_r^2}$$

Note that for any pair of regions (i, j), region j comes before region i in exactly $k!/2$ permutations. Hence, for every (i, j) such that $i \sim j$, $I(j \in N_\pi(i)) = 1$ exactly $k!/2$ times whereas $I(i \in N_\pi(j)) = 1$ for the remaining $k!/2$ permutations and $\frac{1}{k!}\sum_\pi I(j \in N_\pi(i)) = \frac{1}{2}$. Using this we also observe that for any i,

$$\frac{1}{k!}\sum_\pi \sum_{r:i\in N_\pi(r)} \frac{\tau_r}{m_r^2} = \sum_{r\sim i} \frac{\tau_r}{m_r^2}\frac{1}{k!}\sum_\pi I(i \in N_\pi(r)) = \frac{1}{2}\sum_{r\sim i}\frac{\tau_r}{m_r^2}$$

Similarly if r is a common neighbor of i and j, then r is a common directed neighbor of i and j for exactly $k!/3$ permutations, so $\frac{1}{k!}\sum_\pi I(\{i, j\} \subset N_\pi(r)) = \frac{1}{3}$ and consequently

$$\frac{1}{k!}\sum_\pi \sum_{r:\{i,j\}\subset N_\pi(r)} \frac{\tau_r}{m_r^2} = \sum_{r\sim(i,j)} \frac{\tau_r}{m_r^2}\frac{1}{k!}\sum_\pi I(\{i, j\} \subset N_\pi(r)) = \frac{1}{3}\sum_{r\sim(i,j)}\frac{\tau_r}{m_r^2}$$

This proves part (a). The proof for part (b) is very similar to the proof of sparsity for NNGP precision matrices derived in Chapter 2. The $(i, j)^{th}$ element of Q is non-zero if $i \sim j$ or $i \approx j$ or both. If the maximum degree of the graph is m, there can be at most $1 + m + m^2$ non-zero elements in each row. □

Theorem 2 is a powerful result as it demonstrates that, under very general simultaneous specification as in (4.6), we can create a sparse precision matrix over a graph using only the adjacency structure. However, the model still has $2k$ unknown parameters $\{\tau_1, \tau_2, \ldots, \tau_k, m_1, m_2, \ldots, m_k\}$. We need to impose some additional simplifying assumption to solve this final problem. It is tempting to assume homoscedastic errors for the ϵ_i's defined in (4.6). However, when embedded in a hierarchical setup as in (4.3) with Gaussian response, this will lead to identifiability issues between the response noise and ϵ_i's. This also issue arises for the SAR model, which specifies homoscedastic errors thereby limiting its use to only areal generalized linear models for non-Gaussian responses (see Banerjee et al., 2014, for more details on this). Instead, we preserve the heteroscedasticty and, akin to the CAR model, choose $m_i = n_i$ and $\tau_i = n_i\tau_w$ i.e. the number of neighbors for the vertex i. Under these choices of m_i and τ_i, we observe from Theorem 2 that $\Lambda = \tau_w Q_{CAR}$. and $R = \tau_w R^*$ where

$$\begin{aligned} R^*_{ii} &= \tfrac{1}{2}\textstyle\sum_{r\sim i}\frac{1}{n_r} \\ R^*_{ij} &= I(i \approx j)\left(\tfrac{1}{3}\textstyle\sum_{r\sim(i,j)}\frac{1}{n_r}\right) \end{aligned} \tag{4.11}$$

For realizations w over any undirected graph, we can now model w as

$$w \sim N(0, \tau_w Q_{DAGAR}) \tag{4.12}$$

where $Q_{DAGAR} = Q_{CAR} + R^*$. We refer to this as the *Directed Acyclic Graph Autoregressive (DAGAR) model*.

Note that, unlike the proper CAR and SAR, DAGAR does not involve any additional parameters other than the marginal precision τ_w. Any vertex j contributes $\frac{1}{2n_j}$ to $R^*_{i,i}$ if $i \sim j$. Since j has n_j neighbors, it contributes a total of $1/2$ to the trace of R^*, so $trace(R^*) = k/2$. Hence, on average the eigen values of Q_{DAGAR} exceed the eigen values of Q_{CAR} by $1/2$. Most importantly, DAGAR, although constructed using directed acyclic graphs, sheds the dependence on the ordering of the datapoints and still has concise expressions for its entries. However, it is clear from the structure of R^* in (4.11) that, unlike CAR, DAGAR doesn't possess the first order Markovian property although it is second order Markovian. Consequently, the precision matrix of the CAR is sparser and hence implementing CAR models will be faster. Nonetheless, part (b) of Theorem 2 shows that as long as the maximum degree of the graph is reasonably small, Q_{DAGAR} is also very sparse for large k. Simulation experiments detailed in Section 4.5 reveal that under many scenarios, this slight decrease of sparsity for the DAGAR model is often offset by superior performance compared to the CAR model.

4.4 Hierarchical DAGAR models

The DAGAR model is easily embedded into hierarchical setups,like CAR models are. In this section we provide a Gibbs' sampler algorithm for the following hierarchical spatial random intercept model for areal datasets:

$$\begin{aligned} y &= X\beta + w + \epsilon \\ w &\sim N(0, \tau_w Q_{DAGAR}) \\ \epsilon &\sim N(0, \tau_e I_k) \end{aligned} \tag{4.13}$$

Similar to the CAR model, a normal/flat prior for β and inverse-Gamma (IG) priors for τ_e and τ_w provides conjugacy in the Gibbs' sampler. The full hierarchical likelihood

is specified as follows:

$$N(y \mid X\beta + w, \tau_e I_k) \times N(w \mid 0, \tau_w Q_{DAGAR}) \times N(\beta \mid \mu_\beta, Q_\beta) \\ \times IG(\tau_e \mid a_e, b_e) \times IG(\tau_w \mid a_w, b_w) \tag{4.14}$$

where μ_β, Q_β, a_e, b_e, a_w and b_w are known hyper-parameters. One of the main reasons for the popularity of CAR models (or Gaussian Markov Random fields in general) is that the full conditionals for the w_i's are readily available. This along with the sparsity of Q_{CAR} facilitates an efficient Gibbs' sampler. From part (a) of Theorem 2 we see that Q_{DAGAR} is denser than Q_{CAR} as it has non-zero a $(i,j)^{th}$ entry even when i and j are second order neighbors. However, part (b) of the theorem shows that under mild regularity conditions Q_{DAGAR} is very sparse for large k. Also, the full conditionals for w in a DAGAR model is given by: $w_i \mid w_{-i} \sim N(-\frac{1}{Q_{DAGAR,ii}} \sum_{j \neq i} Q_{DAGAR,ij} w_j, \tau_w Q_{DAGAR,ii})$. As the sum in the conditional mean is a sparse sum available in closed form, like in a CAR model, the Gibbs' sampler for DAGAR model is also very efficient. Chapter 2 showed that the number of flops per iteration of a Gibbs sampler in a Nearest Neighbor Gaussian Process model with k locations is $O(km^3)$. The same result holds for DAGAR models for areal datasets with m as the highest degree of the graph, as DAGAR model will have the same sparsity pattern.

4.5 Simulation experiments

We conducted several simulation experiments to assess the performance of DAGAR and CAR. We consider three different graph structures.

4.5.1 One-dimensional path

We considered a simple path graph with $k = 100$ points represented as $1, 2, \ldots, k$, analogous to a time-series. Figure 4.2 shows the structure of a path graph with 10 points. For each point $1 < i < k$, its neighbors are the two points on either side of it. The only neighbor of point 1 is point 2 and the only neighbor of point k is point $k-1$. To generate data on this graph, we embed this graph on the real line where the i^{th} vertex is mapped to i for $i = 1, 2, \ldots, k$. We then generate the spatial random effect vector w from a Matérn GP with precision τ_w, smoothness ν and spatial decay

Figure 4.2: Path graph with 10 vertices

ϕ. We varied the smoothness of the GP ν from 0.5 to 3.5 in increments of one. The spatial decay ϕ was varied over $0.1, 0.2, 0.5$, and 1. Increasing ν represents smoother functions whereas increased ϕ implied shorter spatial range and increasing irregularity (which also means less smoothness). For these choices of ν and ϕ, the average number of 'significant neighbors' (sn) — number of points with correlation greater than 0.5 — varies between 4 (extremely rough) and 89 (extremely smooth).

Subsequently, we generated the response y as $y = w + \delta$ where $\delta \overset{\text{iid}}{\sim} N(0, \tau_e)$. Let $r = \tau_w/\tau_e$ denote the ratio of noise to signal variance. We varied r from 0.001 to 100 thereby covering a very wide spectrum of scenarios. We choose $\log_{10}(r)$ from 16 equally spaced numbers in $[-3, 2]$ to achieve this range for r. For each combination of r, ν and ϕ, we generated 100 datasets and estimated the random effects w using both CAR and $DAGAR$ priors. For comparison, we also fitted the 'oracle' GP which assumed the true ϕ and ν. Estimation was done by maximizing the joint likelihood $p(y, w)$ with respect to w and r. We then obtained the average MSE numbers $||w - \hat{w}||_2^2$ as a metric for model evaluation. Figure 4.3 plots the MSE as a function of r in the log-log-scale for each combination of ν and ϕ. We see that in general the MSE for all three models decreased with decreasing r, which is expected as fitting improves if the data has less noise. However, we see that for small values of r, the DAGAR significantly outperforms CAR. This poor performance of CAR at the left hand side of each plot is not surprising, as small r implies that the relative signal strength is high and CAR models, which tend to oversmooth, cannot estimate such a strong signal. The disparity between the DAGAR and CAR is more prominent for values of ν and ϕ which are tied to increased roughness for the GP surface (figures near the top left of the panel). On the other extreme, when the generating GP is extremely smooth (figures near the bottom left of the panel), DAGAR does perform worse than the CAR. Across most scenarios, the MSE curves for the DAGAR closely resemble those of the oracle GP.

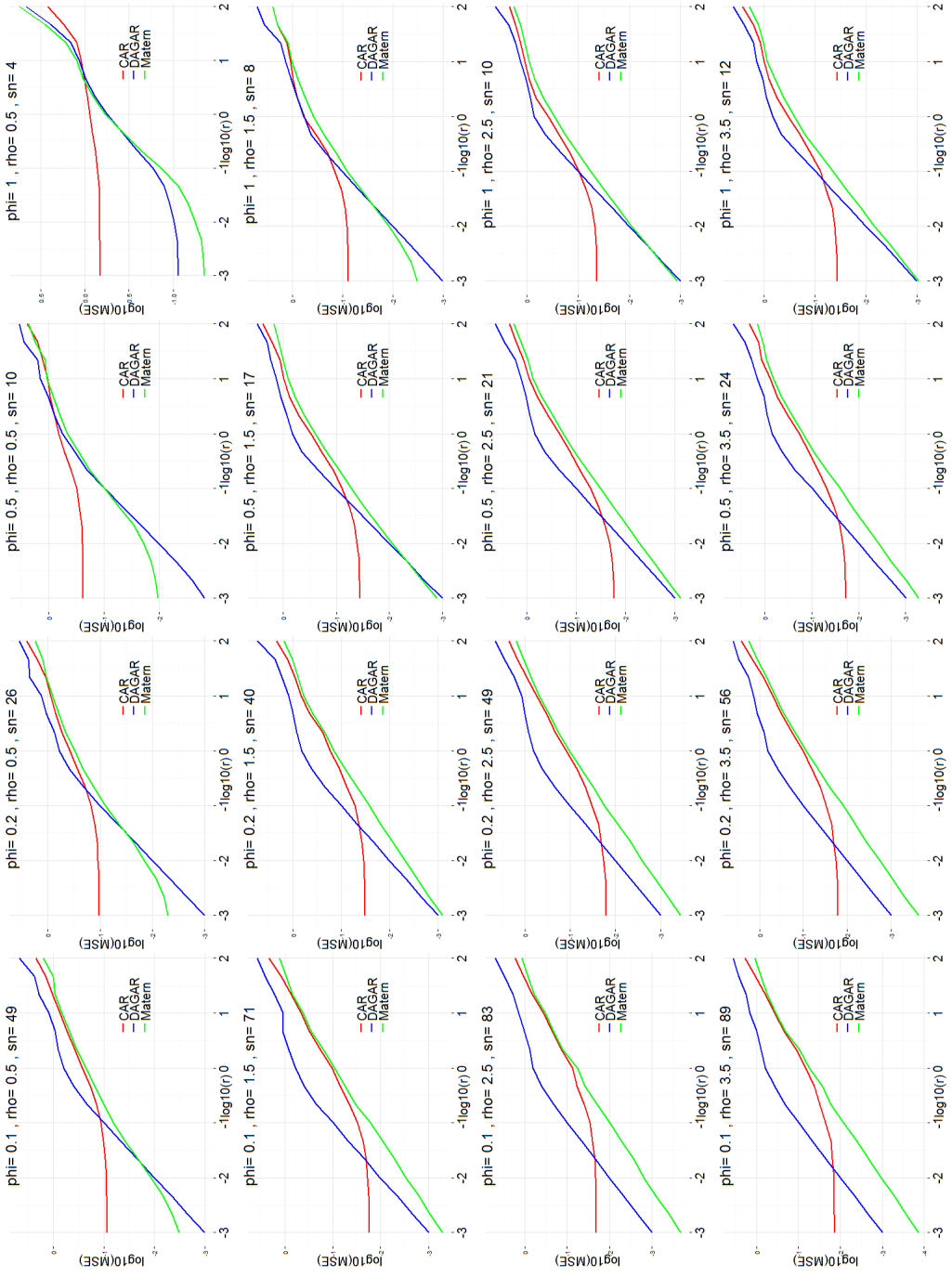


Figure 4.3: Average MSE numbers for path graph

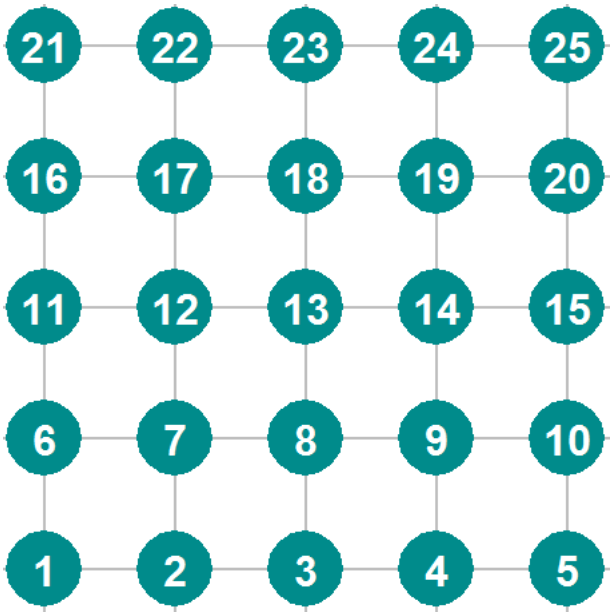


Figure 4.4: Lattice graph with 25 vertices

4.5.2 Two-dimensional Lattice

The second graph we considered was the two-dimensional $m \times m$ lattice or grid graph. The adjacency structure for a sample 5×5 lattice graph is depicted in Figure 4.4. Each interior point has four neighbors — to its north, south, east and west. Each vertex on the edges has three neighbors and the four corner vertices have two neighbors each. For the simulations, we used $m = 10$ so that the total number of vertices is 100 as in Section 4.5.1. We embed the lattice in a two-dimensional plane where each vertex is mapped to the points of a two-dimensional 10×10 grid. The area of the grid is scaled so that the average Euclidean distance between the vertices for the lattice graph is similar to that for the path graph used in Section 4.5.1. We now generate random effects w from a Matérn GP and subsequently generate the response y as in Section 4.5.1. All sets of values for r, ϕ and ν were kept same. Figure 4.5 plots the MSE averaged over 100 replicates. One thing to note here is that the lattice graph is significantly more connected than the path graph. The average number of neighbors for the lattice graph is 3.6 compared to 1.99 for the path graph. Nonetheless, we see that the general trends in Figure 4.3 for the path graphs are reproduced for the lattice graph. For this more

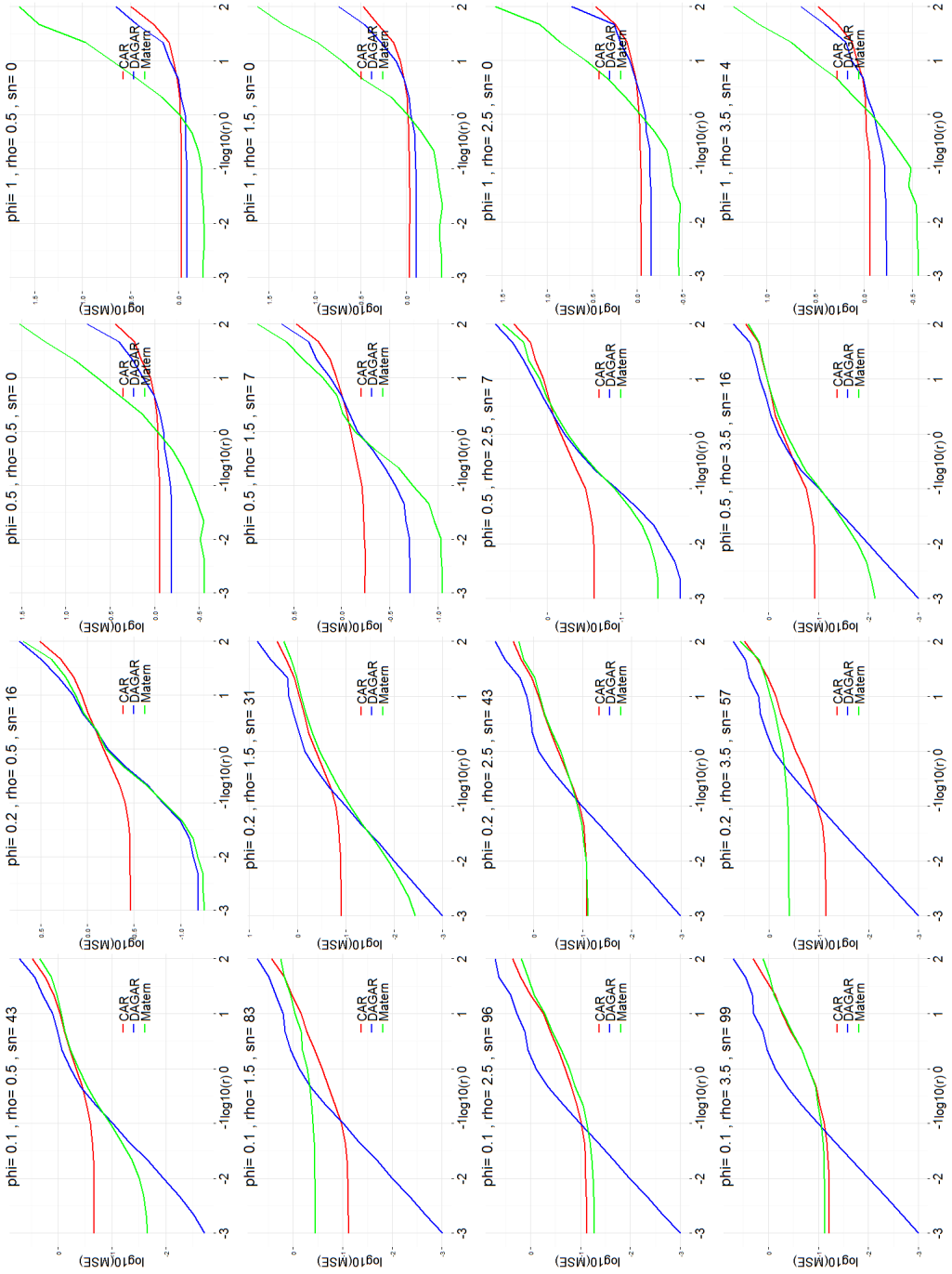


Figure 4.5: Average MSE numbers for lattice graph

connected graph, the disparities between the DAGAR and CAR are even more acute. Even for smoother realizations of the GP (high ν and low ϕ) DAGAR produces lower MSE for a larger range of the noise to signal ratio r. A surprising observation is that for the lattice graph, the oracle GP sometimes performs worse than the DAGAR.

4.5.3 United States State Map

The last graph we considered was the state level map of the contiguous United States. Two states are said to have an edge if they share a common geographical boundary. This creates the adjacency structure for the 48 states depicted in Figure 4.6. The graph has similar connectivity as the lattice graph (the average number of neighbors is 4.5). However, the graph is much more irregular. This is evident from the fact that the standard deviation of the number of neighbors for the USA graph is 1.6 compared to 0.6 for the lattice graph. To embed the graph in a two-dimensional plane, we represent each

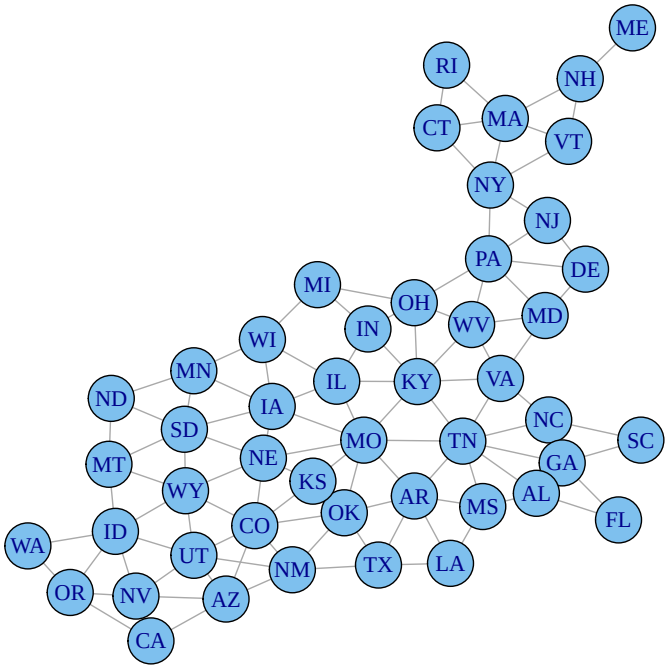


Figure 4.6: Graph for the US states

state by its geographical centroid and use Alber's equal area projections to obtain the mapped two-dimensional co-ordinates representing each state. Note that, this projection is only done to generate the random effects w using Matérn GP on an Euclidean plane. It is not used in fitting the DAGAR or the CAR model, neither of which uses any geographical information beyond the adjacency structure in Figure 4.6. We re-scale the projected points so that the average inter-site distance is similar to the path and grid graph, and then generate w and y using the same settings as in Sections 4.5.1 and 4.5.2. The MSE curves are plotted in Figure 4.7. Once again, the overall trends are similar to those for the grid graph. DAGAR outperforms CAR when the generating GP is rough or the signal strength is high.

4.6 Conclusions

No single model is guaranteed to provide the best fit for every real dataset and it is desirable to have different models tailored to different types of features in the data. The existing repertoire of covariance models used for analyzing areal datasets is extremely limited. In this chapter, we have developed a new class of models for areal datasets that promises to be a significant addition to this inventory. Theoretically, DAGAR models amend several limitations of the CAR model. It ensures that the spatial random effects are endowed with a proper probability distribution and can be used to directly model the response. It does not involve any additional parameters. The precision structure for DAGAR models is sufficiently sparse for large areal datasets thereby facilitating an efficient Gibbs sampler for hierarchical modeling.

Empirically, we have shown that DAGAR models do not suffer from the oversmoothing experienced by CAR models. DAGAR performs significantly better for the datasets on irregular graphs that have strong signals. However, there are concerns about DAGAR overfitting when the true spatial surface is smooth and the noise is high. Further research needs to be conducted to confirm these conjectures and generalize conditions when the DAGAR model will be a better choice. Currently, it may be prudent to fit both CAR and DAGAR models and choose the model which produces better model evaluation score or more realistic estimate of the random effects. Also, setting $m_i = n_i$ and $\tau_i = n_i\tau_w$ is the construction of Q_{DAGAR} is somewhat arbitrary. Perhaps, more

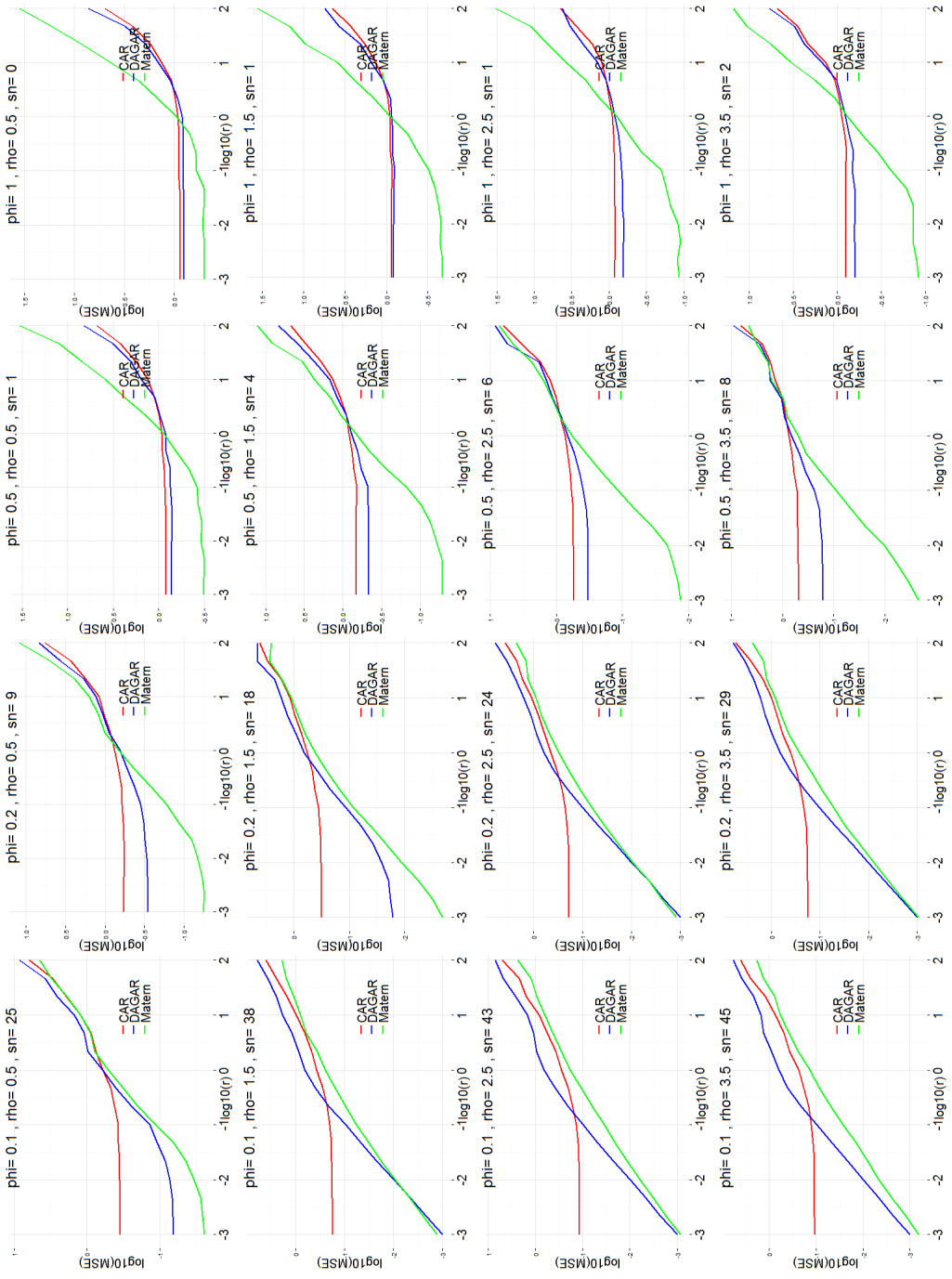


Figure 4.7: Average MSE numbers for USA graph

educated data-based choices for m_i and τ_i may improve DAGAR. However, this would drastically increase the number of parameters in the model. Currently, we restricted our simulation studies to Gaussian responses. As disease data is often observed as counts or proportions, we need to assess the performance of DAGAR models for generalized linear models. We identify these as potential areas of future research.

Chapter 5

CoCoLasso for High-dimensional Error-in-variables Regression

5.1 Introduction

High-dimensional regression has wide applications in various fields such as genomics, finance, medical imaging, climate science, sensor network, etc. The current inventory of high-dimensional regression methods includes Lasso (Tibshirani, 1994), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006) and Dantzig selector (Candès and Tao, 2007) among others. The articles Fan and Li (2006) and Fan and Lv (2010) provide an overview of these existing methods while the book by Buhlmann and van de Geer (2011) discusses their statistical properties in finer details. The canonical high-dimensional linear regression model assumes that the number of available predictors (p) is larger than the sample size (n), although the true number of relevant predictors (s) is much less than n. The model is expressed as $y = X\beta^* + w$ where $y = (y_1, \ldots, y_n)'$ is the vector of responses, $X = ((x_{ij}))$ is the $n \times p$ matrix of covariates, β^* is a $p \times 1$ sparse coefficient vector with only s non-zero entries and $w = (w_1, \ldots, w_n)'$ is the noise vector.

Much of the existing theoretical and applied work on high-dimensional regression has focused on the clean data case. However, we often face corrupted data in many applications where the covariates are observed inaccurately or have missing values. Common

examples include sensor network data (Slijepcevic et al., 2002), high-throughput sequencing (Benjamini and Speed, 2012), and gene expression data (Purdom and Holmes, 2005). It is well known that misleading inference results will be obtained if the regression method for clean data is naively applied to the corrupted data. In order to facilitate further discussion, we assume that we observe a corrupted covariate matrix $Z = (z_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ instead of the true covariate matrix X. Depending on the context, there can be various ways to model the measurement error. In the additive model setup, $z_{ij} = x_{ij} + a_{ij}$ where $A = (a_{ij})$ is the additive error matrix. In the multiplicative error setup, $z_{ij} = x_{ij}m_{ij}$ where m_{ij}s are the multiplicative errors. Missing predictors can be interpreted as a special case of multiplicative measurement errors with $m_{ij} = I(x_{ij}$ is not missing) where $I(\cdot)$ is the indicator function.

Without loss of generality, we take the Lasso as an example to illustrate the impact of measurement errors. We apply the Lasso to the clean data by minimizing:

$$1/(2n)\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{5.1}$$

with respect to β. Here $\lambda > 0$ is the regularization parameter and $\|\cdot\|_p$ denotes the ℓ_p norm for vectors and matrices for $1 \leq p \leq \infty$. If we ignore the measurement error issue, we would apply the Lasso to the corrupted data by minimizing:

$$1/(2n)\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1. \tag{5.2}$$

However, as pointed out in (Rosenbaum and Tsybakov, 2010), the resulting estimate of β is often erroneous if the noise is large. We need to find a proper modification of (5.2) such that its solution is comparable/close to the clean Lasso estimate (5.1).

Observe that the clean Lasso objective function can be equivalently formulated as

$$\frac{1}{2}\beta'\Sigma\beta - \rho'\beta + \lambda\|\beta\|_1 \text{ where } \Sigma = \frac{1}{n}X'X,\ \rho = \frac{1}{n}X'y. \tag{5.3}$$

In (Loh and Wainwright, 2012) Loh and Wainwright use Z and y to construct unbiased surrogates $\widehat{\Sigma}$ for Σ and $\tilde{\rho}$ for ρ. To elucidate, let us consider the classical additive measurement error case. Following (Loh and Wainwright, 2012), assume the additive errors a_{ij} are independent with mean zero and variance τ^2 where τ^2 is a known constant, then

$$E[\frac{1}{n}Z'Z] = \frac{1}{n}X'X + \tau^2\mathbf{I}, \quad E[\frac{1}{n}Z'y - \frac{1}{n}X'y] = 0.$$

Thus Loh and Wainwright suggested using unbiased surrogates

$$\widehat{\Sigma} = \frac{1}{n} Z'Z - \tau^2 \mathbf{I}, \quad \tilde{\rho} = \frac{1}{n} Z'y \tag{5.4}$$

and then solve the following optimization problem to get an estimate of β:

$$\frac{1}{2}\beta'\widehat{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1. \tag{5.5}$$

Although the above solution is very natural, (5.5) is fundamentally different from the clean Lasso. Notice that $\widehat{\Sigma}$ may not be positive semi-definite. When $\widehat{\Sigma}$ does have a negative eigenvalue (which happens very often under high-dimensionality), the objective function in (5.5) is no longer convex. Moreover, the objective function is unbounded from below when $\widehat{\Sigma}$ has a negative eigenvalue. To overcome these technical difficulties, Loh and Wainwright defined their estimator as

$$\hat{\beta} \in \underset{\|\beta\|_1 \leq b_o\sqrt{s}}{\arg\min} \frac{1}{2}\beta'\widehat{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1. \tag{5.6}$$

for some constant b_0. Note that "$\in$" not "$=$" is used in (5.6) because the objective function may still have multiple local/global minimizers even within the region $\|\beta\|_1 \leq b_o\sqrt{s}$. Through some careful analysis, Loh and Wainwright showed that, if b_0 is properly chosen, a projected gradient descent algorithm will converge in polynomial time to a small neighborhood of the set of all global minimizers.

In this article we propose the *Convex Conditioned Lasso (CoCoLasso)* — a convex formulation of the Lasso that can handle a general class of corrupted datasets including the cases of additive or multiplicative measurement error and random missing data. CoCoLasso automatically enjoys the theoretical and computational benefits of convexity that contribute fundamentally to the success of the Lasso. Theoretically, we derive the statistical error bounds of CoCoLasso which are comparable to those given in Loh and Wainwright (2012). Additionally, we establish the asymptotic sign-consistent selection property of CoCoLasso. Earlier Sørensen et al. (2013) derived asymptotic selection consistency properties for the estimator in (5.6) only for the restrictive case of additive measurement error. However, our result does not require any specification of the type of measurement error. This is arguably the most general result for sign consistency in presence of measurement error. There is no sign-consistency result for the non-convex approach by Loh and Wainwright.

Our method has another significant advantage over the non-convex approach by Loh and Wainwright. As mentioned earlier, choosing b_0 in (5.6) is critically important to the estimator by Loh and Wainwright. Their theory requires $b_0 \geq ||\beta^*||_2$ in order to have desirable error bounds. Note that β^* is unknown. On the other hand, b_0 cannot be too large due to the required lower-RE and upper-RE conditions. See Theorem 1 in (5.6) for details. Therefore, in practice one has to carefully choose the b_0 value. Our method does not have this concern. From a pure practical viewpoint, our method uses one tuning parameter λ while the non-convex approach needs two tuning parameters b_0 and λ. CoCoLasso can be readily solved by any efficient algorithm for solving the clean Lasso. For example, we can use the LARS algorithm (Efron et al., 2004) to efficiently compute the entire solution paths for CoCoLasso estimates as λ continuously varies. This is particularly useful for practitioners to understand the procedure.

We notice that in the current literature little attention has been paid to the cross validation methods used for corrupted data. Simply replacing Z by X leads to biased version of the cross validation procedure (similar to (5.5) being a biased version of (5.3)). We demonstrate how the ideas used to develop CoCoLasso can be seamlessly adapted to propose new corrected cross-validation technique tailored for data with measurement error. To our best knowledge, the existing work on high-dimensional regression with measurement error did not touch on this cross-validation issue. The new corrected cross-validation has its own independent importance.

It is worth pointing out that a Dantzig Selector type estimator named matrix uncertainty (MU) estimator was proposed in Rosenbaum and Tsybakov (2010) for additive measurement error models. An improved version of MU estimator was proposed in Rosenbaum and Tsybakov (2013). Belloni et al. (2014b) establishes near-optimal minimax properties of the estimator in Rosenbaum and Tsybakov (2013) and develops a conic-programming based estimator that achieves minimax bounds. Two more conic programming based estimators have been recently proposed in Belloni et al. (2014a) for the same model setup. It has been empirically observed that solving the Lasso problem can be much faster than solving the Dantzig selector Efron et al. (2007). Compared to Dantzig Selector type estimators and the conic programming based estimators, the direct Lasso-modification methods, such as CoCoLasso, would enjoy computational advantages, which is very important for high-dimensional data analysis.

The rest of the article is organized as follows. In Section 5.2 we define the CoCo-Lasso estimator. In Section 5.3 we discuss the main theoretical results. In Section 5.4 we discuss the consequences of the results in Section 5.3 for additive and multiplicative measurement error setups. A new cross-validation technique for corrupted data is developed in Section 5.5. In Section 5.6 we present simulation results to demonstrate the empirical performance of CoCoLasso.

5.2 CoCoLasso

We first introduce some necessary notations and model setup. For any matrix $K = ((k_{ij}))$, we write $K > 0$ (≥ 0) when it is positive (semi-)definite. Let $\|K\|_\infty = \max_i \sum_j |k_{ij}|$ denote the matrix ℓ_∞ norm whereas $\|K\|_{\max} = \max_{i,j} |k_{ij}|$ denote the elementwise maximum norm. Also let $\Lambda_{\min}(K)$ and $\Lambda_{\max}(K)$ denote the minimum and maximum eigen values of K respectively. We assume that all variables are centered so that the intercept term is not included in the model and the covariance matrix X has normalized columns i.e. $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$ for every $j = 1, \ldots, p$. Without loss of generality, assume that $S = \{1, 2, \ldots, s\}$ is the true support set of the regression coefficient vector and write $\beta^* = (\beta_S^{*T}, 0')'$ and $X = (X_S, X_{S^c})$. Hence the true model can be rewritten as $y = X_S\beta_S^* + w$ where the components of β_S^* are non-zero. For any vector v, we can partition it as $v = (v_S', v_{S^c}')'$. Also, we partition Σ as

$$\Sigma = \begin{pmatrix} (1/n)X_S'X_S & (1/n)X_S'X_{S^c} \\ (1/n)X_{S^c}'X_S & (1/n)X_{S^c}'X_{S^c} \end{pmatrix} = \begin{pmatrix} \Sigma_{S,S} & \Sigma_{S,S^c} \\ \Sigma_{S^c,S} & \Sigma_{S^c,S^c} \end{pmatrix}$$

In this work we consider the fixed design case because we want to avoid the identifiability issues between the true design matrix and the measurement error matrix. In the theoretical literature on the clean Lasso, it is often assumed that w_i's are independent and identically distributed sub-Gaussian random variables with parameter σ^2. We use the same assumption here.

As mentioned earlier, in a clean setting where the predictor matrix X is observed accurately, a Lasso estimate is obtained by minimizing (5.3). When the dataset is corrupted by measurement errors, the observed matrix of predictors Z is some function of the true covariance matrix X and random errors. Based on Z and y, estimates $\widehat{\Sigma}$ and $\tilde{\rho}$ are constructed as surrogates to replace Σ and ρ respectively in (5.3). Different

pairs of unbiased estimates $(\widehat{\Sigma}, \tilde{\rho})$ are provided in Loh and Wainwright (2012) for various types of measurement errors. We will present the actual form of $(\widehat{\Sigma}, \tilde{\rho})$ in section 4, but for now we only need to assume that $(\widehat{\Sigma}, \tilde{\rho})$ have been computed.

We now define a nearest positive semi-definite matrix projection operator as follows: for any square matrix K,

$$(K)_+ = \arg\min_{K_1 \geq 0} \|K - K_1\|_{\max}.$$

Then we denote $\widetilde{\Sigma} = (\widehat{\Sigma})_+$ and define our *Convex conditioned Lasso (CoCoLasso)* estimate as

$$\hat{\beta} = \arg\min_{\beta} (1/2)\beta'\widetilde{\Sigma}\beta - \tilde{\rho}'\beta + \lambda\|\beta\|_1 \tag{5.7}$$

We use an alternating direction method of multipliers (ADMM) (Boyd et al., 2011) to obtain $\widetilde{\Sigma}$ from $\widehat{\Sigma}$. The ADMM algorithm is very efficient and details of the algorithm are provided in Appendix 5.9. By definition, $\widetilde{\Sigma}$ is always positive semi-definite. Note that Σ is positive semi-definite when $p > n$. Subsequently, we can reformulate our problem as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}||\widetilde{y} - \widetilde{Z}\beta||_2^2 + \lambda\|\beta\|_1 \tag{5.8}$$

where $\widetilde{Z}/\sqrt{n}$ is the Cholesky factor of $\widetilde{\Sigma}$ i.e. $\frac{1}{n}\widetilde{Z}'\widetilde{Z} = \widetilde{\Sigma}$ and $\widetilde{y}$ is such that $\widetilde{Z}'\widetilde{y} = Z'y$.

Numerically, (5.8) is just like the clean Lasso. One can apply several very fast solvers to solve (5.7), such as the coordinate descent algorithm (Friedman et al., 2010) or the homotopy algorithm (Efron et al., 2004). This is a great advantage for practitioners, as the Lasso solvers are widely used in practice and many know how to use them. We use the LARS-EN algorithm to obtain the solution as it simultaneously provides the entire solution path for different values of λ.

Theoretically, (5.7) can be analyzed by the tools for analyzing the clean Lasso. The surrogate $\widehat{\Sigma}$ chosen by Loh and Wainwright (2012) is often an unbiased estimate of the true gram matrix Σ, achieving a desired rate of convergence under the max norm. By definition, we have

$$\|\widetilde{\Sigma} - \Sigma\|_{\max} \leq \|\widetilde{\Sigma} - \widehat{\Sigma}\|_{\max} + \|\widehat{\Sigma} - \Sigma\|_{\max} \leq 2\|\widehat{\Sigma} - \Sigma\|_{\max} \tag{5.9}$$

Equation (5.9) ensures that $\widetilde{\Sigma}$ approximates Σ as well as the initial surrogate $\widehat{\Sigma}$.

Compared with Loh and Wainwright's estimator in Loh and Wainwright (2012), CoCoLasso is guaranteed to be convex. This avoids the need of doing any non-convex analysis of the method. Furthermore, unlike Loh and Wainwright (2012) our method does not require any knowledge of $\|\beta\|_1$ and thereby eliminates the need for an initial estimate to obtain a bound for $\|\beta\|_1$. In the next section, we show that CoCoLasso is sign consistent and has the ℓ_1, ℓ_2 error bounds comparable to that in Loh and Wainwright (2012).

5.3 Theoretical Analysis

In this section we derive the ℓ_1 and ℓ_2 bounds for the statistical error of the CoCoLasso estimate as well as its support recovery probability bounds.

5.3.1 ℓ_1 and ℓ_2 bounds for the statistical error

We assume that $\widehat{\Sigma}$ and $\tilde{\rho}$ are sufficiently 'close' to Σ and ρ respectively in the following sense:

Definition 1. *Closeness condition: Let us assume that the distribution of $\widehat{\Sigma}$ and $\tilde{\rho}$ are identified by a set of parameters θ. Then there exists universal constants C and c, and positive functions ζ and ϵ_0 depending on β^*_S, θ and σ^2 such that for every $\epsilon \leq \epsilon_0$, $\widehat{\Sigma}$ and $\tilde{\rho}$ satisfy the following probability statements:*

$$\begin{aligned} Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \epsilon) &\leq C \exp\left(-cn\epsilon^2\zeta^{-1}\right) \ \forall\, i, j = 1, \ldots, p \\ Pr(|\tilde{\rho}_j - \rho_j| \geq \epsilon) &\leq C \exp\left(-cns^{-2}\epsilon^2\zeta^{-1}\right) \ \forall\, j = 1, \ldots, p \end{aligned} \tag{5.10}$$

The Closeness Condition requires that the surrogates $\widehat{\Sigma}$ (and hence $\widetilde{\Sigma}$) and $\tilde{\rho}$ are close to Σ and ρ respectively in terms of the elementwise maximum norm. We show later in Section 5.4 that this condition is satisfied by the surrogates defined in Loh and Wainwright (2012) for commonly used additive or multiplicative measurement error models.

We also assume the following compatibility or restricted eigenvalue condition:

$$0 < \Omega = \min_{x \neq 0,\, \|x_{S^c}\|_1 \leq 3\|x_S\|_1} \frac{x'\Sigma x}{\|x\|_2^2} \tag{5.11}$$

Restricted eigenvalue condition similar to this has been used in van de Geer and Buhlmann (2009) to obtain bounds of statistical error of the clean Lasso estimate.

We now state the main result on the statistical error of the CoCoLasso estimate. All proofs are provided in Section 5.8. Note that, for all the theoretical results, C and c denote generic positive constants. Their values vary from expression to expression but they remain universal constants.

Theorem 1. *Under the assumptions (5.10) and (5.11), for* $\lambda \leq min(\epsilon_0, 12\epsilon_0\|\beta^*_S\|_\infty)$ *and* $\epsilon \leq min(\epsilon_0, \Omega/64s)$ *the following results holds true with probability at least* $1 - p^2C\exp\left(-cns^{-2}\lambda^2\zeta^{-1}\right) - p^2C\exp\left(-cn\epsilon^2\zeta^{-1}\right)$*:*

$$\|\hat{\beta} - \beta^*\|_2 \leq C\lambda\sqrt{s}/\Omega \quad , \quad \|\hat{\beta} - \beta^*\|_1 \leq C\lambda s/\Omega \tag{5.12}$$

Results similar to Theorem 1 were derived in Theorems 1 and 2 of Loh and Wainwright (2012) for the estimates obtained by projected gradient descent algorithm for the non-convex objective function. Both the ℓ_1 and ℓ_2 bounds obtained in Theorem 1, are of the same order as the analogous bounds for statistical error of the traditional Lasso estimate. The tail probability depends on the presence of error in the variables through the component ζ. Precise expression for ζ is derived for the case of additive measurement error in Section 5.4.

5.3.2 Sign consistency

In order to establish the sign consistency of CoCoLasso, in addition to the closeness conditions in (5.10), we assume the irrepresentable and minimum eigenvalue conditions on Σ which are sufficient and nearly necessary for sign consistency of the clean Lasso (Zou, 2006; Zhao and Yu, 2006; Wainwright, 2009):

$$\|\Sigma_{S^c,S}\Sigma^{-1}_{S,S}\|_\infty = 1 - \gamma < 1, \qquad \Lambda_{\min}(\Sigma_{S,S}) = C_{\min} > 0 \tag{5.13}$$

The main result on recovery of signed support is stated as follows:

Theorem 2. *Under the assumptions given in Equations (5.10) and (5.13), for* $\lambda \leq min(\epsilon_0, 4\epsilon_0/\gamma)$ *and* $\epsilon \leq \min(\epsilon_1, \lambda/(\lambda\epsilon_2 + \epsilon_3))$ *where* ϵ_i*'s are bounded positive constants depending of* $\Sigma_{S,S}$*,* β^*_S*,* θ *and* σ^2*, the following occurs with probability at least* $1 - \delta_1$ *where* $\delta_1 = p^2C\exp\left(-cns^{-2}\gamma^2\lambda^2\zeta^{-1}\right) + p^2C\exp\left(-cns^{-2}\epsilon^2\zeta^{-1}\right)$

(a) There exists a unique solution $\hat{\beta}$ minimizing (5.7) whose support is a subset of the true support.

(b) $||\hat{\beta}_S - \beta^*_S||_\infty \leq \kappa\lambda$ *where* $\kappa = \left(4||\Sigma^{-1}_{S,S}||_\infty + C^{-1/2}_{\min}\right)$

(c) If $|\beta^*_{\min}| \geq \kappa\lambda$, *then sign*$(\hat{\beta}_S)$ =*sign*(β^*_S)

If we assume for simplicity that κ is $\mathcal{O}(1)$ and the triplet $\{n, p, s\}$ and β^* satisfy the scaling:

$$\begin{aligned} s^2 \log p/n \to 0 \text{ as } n, p \to \infty \\ |\beta^*_{\min}| \gg s(\zeta \log p/n)^{1/2} \end{aligned} \tag{5.14}$$

then from the expression of δ_1 in Theorem 2 we can choose λ so that $1 - \delta_1$ goes to one, which implies the sign-consistency of the CoCoLasso estimate.

Corollary 1. *If* Σ, $\widetilde{\Sigma}$ *and* $\tilde{\rho}$ *satisfy the regularity conditions given in Theorem 2, then under the scaling in Equation (5.14), the CoCoLasso estimate* $\hat{\beta}$ *defined in (5.7) is sign-consistent if* $|\beta^*_{\min}| \gg \lambda \gg s(\zeta \log p/n)^{1/2}$ *and we also have the* ℓ_∞ *error bound* $Pr(||\hat{\beta}_S - \beta^*_S||_\infty \leq \kappa\lambda) \to 1$.

So far in this section we have derived a general theory for the CoCoLasso where there is no assumption on the type of measurement error and the form of the estimates $\widehat{\Sigma}$ and $\tilde{\rho}$. The only condition that requires a careful check is that the estimates $\widehat{\Sigma}$ and $\tilde{\rho}$ are close enough to Σ and ρ respectively in the sense defined in (5.10). In the next section, we consider two specific types of error-in-variables models and use the results of this section to derive the theoretical properties of CoCoLasso estimates for those models.

5.4 CoCoLasso under Two Types of Measurement Errors

5.4.1 Additive error

We assume that the entries of the observed design matrix Z is contaminated by additive measurement error i.e. $z_{ij} = x_{ij} + a_{ij}$ or in matrix notation, $Z = X + A$ where $A = ((a_{ij}))$ is the matrix of measurement errors. We also assume that the rows of A are independent and identically distributed with 0 mean, finite covariance Σ_A and sub-Gaussian parameter τ^2. Following (Loh and Wainwright, 2012) we assume that Σ_A is known. The unbiased estimates of Σ and ρ are given by $\widehat{\Sigma}_{add} = \frac{1}{n}Z'Z - \Sigma_A$ and

$\tilde{\rho}_{add} = \frac{1}{n}Z'y$, respectively. It is easy to observe that $\widehat{\Sigma}_{add}$ can have negative eigenvalues precluding convex optimization. CoCoLasso estimates for this model will be based on the modified objective function

$$\tilde{f}_{add}(\beta) = (1/2)\beta'\widetilde{\Sigma}_{add}\beta - \tilde{\rho}'_{add}\beta + \lambda|\beta\|_1 \text{ where } \widetilde{\Sigma}_{add} = (\widehat{\Sigma}_{add})_+.$$

The following results show that $\widehat{\Sigma}_{add}$ and $\tilde{\rho}_{add}$ satisfy the conditions in Equation (5.10).

Lemma 1. *$\widehat{\Sigma}_{add}$ and $\tilde{\rho}_{add}$ satisfy the closeness conditions in (5.10) with $\epsilon_0 = c\tau^2$ and $\zeta = max(\tau^2, \tau^2\|\beta^*_S\|^2_\infty, \tau^4, \sigma^4)$.*

So, even though $\widehat{\Sigma}_{add}$ may not be positive definite, the surrogates $\widehat{\Sigma}_{add}$ and $\tilde{\rho}$ satisfy (5.10). The following result is an immediate consequence:

Corollary 2. *The results of Theorems 1 and 2 (and Corollary 1) hold for the CoCoLasso estimate for the additive error model under the assumptions (5.11) and (5.13) (and (5.14)) respectively.*

As ζ increases with τ we see that the lower bound for λ required in the Corollary increases as τ increases implying that more penalization is required in presence of larger measurement error to accurately recover the sparse support.

Note that the additive error covariance Σ_A is assumed to be known in order to compute the CoCoLasso estimate. Similar assumption was used in Loh and Wainwright (2012) and Rosenbaum and Tsybakov (2013) as it is unclear how to obtain a data-driven estimate of Σ_A when only one dataset is available. If however, multiple replicates of the data are available, following Loh and Wainwright (2012), one can obtain a data-driven estimate $\widehat{\Sigma}_A$ of Σ_A and define $\widehat{\Sigma}_{add} = \frac{1}{n}Z'Z - \widehat{\Sigma}_A$.

5.4.2 Multiplictive error and missing data

If we assume that the errors are multiplicative, we observe $z_{ij} = x_{ij}m_{ij}$. In matrix notation, we have $Z = X \odot M$ where $M = ((m_{ij}))$ and $\odot$ denotes the elementwise multiplication operator for vectors and matrices. We assume that the rows of M are independent and identically distributed with mean μ_M, covariance Σ_M and sub-Gaussian parameter τ^2. Under the assumption that the entries of μ_M and $\Sigma_M + \mu_M\mu'_M$ are

strictly positive, Loh and Wainwright (2012) suggests using the unbiased surrogates $\widehat{\Sigma}_{mult} = (1/n)ZZ' \oslash (\Sigma_M + \mu_M\mu_M')$ and $\tilde{\rho}_{mult} = (1/n)Z'y \oslash \mu_M$ where $\oslash$ denotes the elementwise division operator for vectors and matrices. $\widehat{\Sigma}_{mult}$ once again may not be positive semi-definite. The CoCoLasso estimate $\hat{\beta}$ is obtained as

$$\min_{\beta}(1/2)\beta'(\widetilde{\Sigma}_{mult})_+\beta - \tilde{\rho}_{mult}'\beta + \lambda|\beta\|_1 \text{ where } \widetilde{\Sigma}_{mult} = (\widehat{\Sigma}_{mult})_+.$$

Randomly missing covariates can be formulated as a multiplicative error model. For example, a simple model assumes that x_{ij}'s are missing randomly with probability r and their missing statuses are independent of one another. Then we can defining $z_{ij} = x_{ij}m_{ij}$ where $m_{ij} = I(x_{ij} \text{ is not missing }) \sim Bernoulli(1-r)$. Other missing data models with different choices of the missing probabilities (e.g. $m_{ij} \sim Bernoulli(1-r_j)$) will also fall under the same setup. We can obtain estimate of r (or r_j) as the proportion of missing entries in the matrix (or in the j^{th} column). For simplicity, we can assume r is known and then Σ_M and μ_M are known as well.

We now establish analogous results for the CoCoLasso estimate in this multiplicative model setup. Note that as the errors are multiplicative, in order to have all the z_{ij}'s to be close to the respective x_{ij}'s, we need an upper bound for both x_{ij} and m_{ij}. We also need a positive lower bound for the entries of μ_M and $\Sigma_M + \mu_M\mu_M'$ for the expressions of $\widehat{\Sigma}_{mult}$ and $\tilde{\rho}_{mult}$ to be meaningful. To ensure these, we impose the following additional set of regularity conditions for the multiplicative setup:

$$\begin{aligned} \max_{i,j} |X_{ij}| = X_{\max} < \infty, \qquad & \min_{i,j} E(m_1m_1') = M_{\min} > 0 \\ \min\ \mu_M = \mu_{\min} > 0, \qquad & \max\ \mu_M = \mu_{\max} < \infty \end{aligned} \tag{5.15}$$

Under these regularity conditions the following lemma shows that $\widetilde{\Sigma}_{mult}$ and $\tilde{\rho}_{mult}$ satisfies the conditions in (5.10):

Lemma 2. *There exists positive functions ϵ_0 and ζ depending on β_S^*, τ^2, σ^2 and the constants in (5.15) such that $\widehat{\Sigma}_{mult}$ and $\tilde{\rho}_{mult}$ satisfy the closeness conditions in (5.10).*

Having proved Lemma 2, once again we use Theorems 1, 2 and Corollary 1 to have the following results:

Corollary 3. *The results of Theorems 1 and 2 (and Corollary 1) hold for the CoCoLasso estimate for the multiplicative error/missing data model under the assumptions (5.11) and (5.13) (and (5.14)) respectively.*

5.5 Corrected cross-validation

In applications, cross-validation (Hastie et al., 2011) is a widely used technique for choosing the tuning parameter in penalized methods. However, cross validation for data corrupted with measurement error has received very little attention. In the presence of noisy/corrupted data, naive application of cross-validation is biased and a novel correction is needed. To elucidate, consider the usual K-fold cross validation for selecting the tuning parameter in the clean Lasso. Let $(X_k,\, y_k)$ denote the true design matrix and response vector for the k^{th} fold of the data for $k = 1, 2, \ldots, K$. Likewise, let $(X_{-k},\, y_{-k})$ denote the design matrix and response vector respectively after removing the k^{th} fold. In absence of measurement error, the estimate for the prediction error for the k^{th} fold is given by $err_k(\lambda) = \frac{1}{n_k}\|y_k - X_k\hat{\beta}_k(\lambda)\|_2^2$ where n_k is the size of the k^{th} fold and $\hat{\beta}_k(\lambda)$ is the Lasso estimate based on X_{-k}, y_{-k} with tuning parameter λ. The optimal λ is obtained by minimizing the total cross-validation error, i.e.,

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\|y_k - X_k\hat{\beta}_k(\lambda)\|_2^2. \tag{5.16}$$

However, when we face noisy/corrupted data, as X is unknown or partially missing, (5.16) is not directly available. If we naively use the observed data (Z, y), then the cross-validated choice of λ is defined by minimizing

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\|y_k - Z_k\hat{\beta}_k(\lambda)\|_2^2. \tag{5.17}$$

Even when we use the CoCoLasso (or the estimator in 5.6) to compute $\hat{\beta}_k(\lambda)$ based on Z_{-k}, y_{-k}, the above criterion is biased compared to (5.16) in the same way the loss function in (5.5) is a biased version of (5.3).

Using simple algebra we observe that (5.16) is equivalent to

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k(\lambda)'\Sigma_k\hat{\beta}_k(\lambda) - 2\rho_k'\hat{\beta}_k(\lambda), \tag{5.18}$$

where $\Sigma_k = \frac{1}{n_k}X_k'X_k$ and $\rho_k = \frac{1}{n_k}X_k'y_k$.

It may seem that using the unbiased surrogates $\widehat{\Sigma}_k$ and $\tilde{\rho}_k$ in (5.18) may overcome the bias issue. However, as $\widehat{\Sigma}_k$ possibly has negative eigen values this will lead to a cross validation function unbounded from below.

In the light of the above discussion, we propose a new cross validation method for corrupted data that adapts the same central idea used to construct CoCoLasso i.e. we can use $(\widehat{\Sigma}_k)_+$ and $\tilde{\rho}_k$ in (5.18). With this correction, the cross-validated λ is defined as

$$\tilde{\lambda} = \arg\min_{\lambda} \sum_{k=1}^{K} \hat{\beta}_k(\lambda)'(\widehat{\Sigma}_k)_+\hat{\beta}_k(\lambda) - 2\tilde{\rho}_k'\hat{\beta}_k(\lambda). \tag{5.19}$$

We call the above procedure the corrected cross-validation.

5.6 Numerical Studies

We use simulated datasets to evaluate the performance of CoCoLasso. For comparison we also included the Loh and Wainwright's method described in Loh and Wainwright (2012). For convenience, we use NCL (Non-convex Lasso) to denote Loh and Wainwright's method in this section.

5.6.1 Simulation Models

We considered both additive measurement errors and multiplicative measurement errors in the simulation study.

Additive errors case. We generate data from the model $y \sim N(X\beta^*, \sigma^2 I)$ where

$$\beta^* = (3, 1.5, 0, 0, 2, 0, \ldots, 0)'.$$

The sample size n is set to be 100 and $p = 250$. The rows of X are independent and identically distributed normal random variables with mean zero and covariance matrix Σ_X. We consider two models for Σ_X — autoregressive ($\Sigma_{X,ij} = 0.5^{|i-j|}$) and compound symmetry ($\Sigma_{X,ij} = 0.5 + I(i = j) * 0.5$). We set $\sigma = 3$ giving a signal to noise ratio of 2.36 for autoregressive (AR) and 3.20 for compound symmetry (CS). We generate $Z = X + A$ where the rows of A are independent and identically distributed $N(0, \tau^2 I)$ where $\tau = 0.75$, 1 and 1.25.

Multiplicative Errors case. We also evaluated the performance of CoCoLasso and NCL in a multiplicative errors setup. The true model is assumed to be same as

in the additive error setup. We now generate $Z = X \odot M$ where we assume that the elements of $M = ((m_{ij}))$ follow log-normal distribution i.e. $\log(m_{ij})$'s are independent and identically distributed $N(0, \tau^2)$ where $\tau = 0.25$, 0.5 and 0.75.

5.6.2 Simulation results and conclusions

We used 5-fold corrected cross-validation for the CoCoLasso in our numerical examples. The code for NCL was provided by Dr. Po-Ling Loh. NCL requires an initial estimator. Following Sørensen et al. (2013), the initial estimate is a naive Lasso estimate based on y and Z which is tuned by 5-fold cross validation. NCL also requires knowledge of $\|\beta_S^*\|_1$ for choosing the constraint parameter. Since, this is impossible to know beforehand, a naive 5-fold cross validation was used to select the optimal R from 100 equally spaced values in $[R_{max}/500, 2 * R_{max}]$ where R_{max} is the ℓ_1 norm of the initial estimate.

The accuracy of estimators is gauged by the Prediction Error (PE) and the Mean Squared Error (MSE) where

$$PE(\hat{\beta}) = (\beta^* - \hat{\beta})' \Sigma_X (\beta^* - \hat{\beta})$$

and

$$MSE(\hat{\beta}) = \|\beta^* - \hat{\beta}\|_2^2.$$

To evaluate variable selection, we record C and IC that denote the number of correct and incorrect predictors identified, respectively.

Table 5.1 and Table 5.2 summarize the simulation results for the additive error case and the multiplicative error case, respectively. We observe that CoCoLasso is more accurate than NCL as measured by PE and MSE, and the gap between the two methods widens as the perturbation level increases (measured by τ). NCL tends to select a sparser model than CoCoLasso, it tends to miss importance variables as the noise level is high.

5.7 Summary

In this paper we have proposed a novel convex approach to modify the classical Lasso with the clean data to handle the noisy data case. Our approach, named CoCoLasso, is easy to understand, easy to use and has solid theoretical foundations. We also have

Table 5.1: Summary statistics for the additive error simulation study based on 100 replications. Reported numbers are the medians and standard errors (se) are computed by bootstrap. "CoCo" stands for CoCoLasso. "NCL" is the method in Loh and Wainwright (2012). AR denotes Autoregressive covariance for the predictors whereas CS denotes compound symmetry covariance.

		$\tau = 0.75$		$\tau = 1.0$		$\tau = 1.25$	
		CoCo	NCL	CoCo	NCL	CoCo	NCL
	C	3	3	3	2	3	2
	IC	11	3	11	1	10	0
AR	PE	3.66	4.13	5.8	6.91	8.49	10.92
	se(PE)	0.19	0.26	0.26	0.34	0.5	0.46
	MSE	3.81	3.76	5.57	6.07	7.94	8.36
	se(MSE)	0.19	0.18	0.2	0.27	0.24	0.3
	C	2	2	2	1.5	2	1
	IC	14	11.5	18	7	21	5
CS	PE	4.49	4.57	6.03	6.91	6.99	10.47
	se(PE)	0.22	0.31	0.22	0.34	0.25	0.58
	MSE	8.05	8.03	11.01	10.31	12.97	15.06
	se(MSE)	0.33	0.48	0.37	0.99	0.34	1.06

devised a novel cross validation methods for corrupted data. We have demonstrated the superior performance of our method over the non-convex approach in Loh and Wainwright (2012) by simulation studies.

Finally, we would like to comment on the generality of the CoCoLasso approach. Although we use the Lasso to illustrate the idea of CoCoLasso, the basic approach of CoCoLasso can be directly used in conjunction with other popular convex penalized methods. For example, the fused Lasso Tibshirani et al. (2005) is a popular technique for ordered variable selection. Following the development of CoCoLasso, we can readily develop CoCo-FusedLasso. We opt not to discuss these variants in the present paper.

Table 5.2: Summary statistics for the multiplicative error simulation study based on 100 replications. Reported numbers are the medians and standard errors (se) are computed by bootstrap. "CoCo" stands for CoCoLasso. "NCL" is the method in Loh and Wainwright (2012). AR denotes Autoregressive covariance for the predictors whereas CS denotes compound symmetry covariance.

		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		CoCo	NCL	CoCo	NCL	CoCo	NCL
	C	3	3	3	3	3	2
	IC	14	12	12	6	10	1
AR	PE	2.02	2.47	3.25	3.58	7.32	8.32
	se(PE)	0.15	0.18	0.14	0.25	0.2	0.29
	MSE	1.95	2.26	2.93	3.09	6.19	6.58
	se(MSE)	0.09	0.14	0.14	0.18	0.2	0.26
	C	3	3	3	3	2	1
	IC	15	18	13	11	16	4
CS	PE	2.23	2.37	3.66	3.82	7.93	9.31
	se(PE)	0.16	0.1	0.15	0.19	0.3	0.41
	MSE	4.21	4.32	6.11	5.75	10.43	9.34
	se(MSE)	0.27	0.21	0.27	0.26	0.25	0.61

5.8 Proofs

In this section we present the proofs of Theorems 1 and 2 as well as Lemmas 1 and 2. A few useful properties and technical results about sub-Gaussian random variables required in the proofs are provided in Appendix 5.10. Throughout this section we denote C and c to be universal constants whose values may vary across different expressions. We also introduce a few additional notations used subsequently in the proofs.

$$\begin{array}{cccc} D = \widetilde{\Sigma} - \Sigma, & G = \Sigma_{S^c,S}\Sigma_{S,S}^{-1}, & \widetilde{G} = \widetilde{\Sigma}_{S^c,S}\widetilde{\Sigma}_{S,S}^{-1}, & H = \widetilde{G} - G \\ F = \widetilde{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}, & \phi = ||\Sigma_{S,S}^{-1}||_\infty, & \psi = ||\Sigma_{S,S}||_\infty, & B = ||\beta_S^*||_\infty \end{array} \tag{5.20}$$

5.8.1 Proof of Theorem 1

We first state and prove a simple result which will be later used in the proof:

Lemma 3. *For any $\epsilon > 0$ we have,*

$$Pr(\|\widetilde{\Sigma} - \Sigma\|_{\max} \geq \epsilon) \leq p^2 \max_{i,j} Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \epsilon/2) \tag{5.21}$$

Proof. From Equation (5.9) we have

$$Pr(\|\widetilde{\Sigma} - \Sigma\|_{\max} \geq \epsilon) \leq Pr(\|\widehat{\Sigma} - \Sigma\|_{\max} \geq \epsilon/2)$$

The proof then follows using union bounds over $Pr(|\widehat{\Sigma}_{ij} - \Sigma_{ij}| \geq \epsilon/2)$. □

Proof of Theorem 1. The general idea of the proof of this Theorem closely resembles the proofs of (Buhlmann and van de Geer, 2011, Lemma 6.3 and Theorem 6.1) for obtaining the error bounds of the traditional Lasso estimate. From the definition of $\hat{\beta}$ in (5.7), we have

$$\frac{1}{2}\hat{\beta}'\widetilde{\Sigma}\hat{\beta} - \tilde{\rho}'\hat{\beta} + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\beta^{*T}\widetilde{\Sigma}\beta^* - \tilde{\rho}'\beta^* + \lambda\|\beta^*\|_1$$

Expanding $\hat{\beta}$ as $\hat{v} + \beta^*$ where $\hat{v} = \hat{\beta} - \beta^*$, this simplifies to

$$\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}\|_1 \leq \hat{v}'(\tilde{\rho} - \widetilde{\Sigma}\beta^*) + \lambda\|\beta^*\|_1 \leq ||\hat{v}\|_1||\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty + \lambda\|\beta^*\|_1 \tag{5.22}$$

In order to obtain an upper bound for the left hand side we first bound the quantity $\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty$. Using triangular inequality we have

$$\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty \leq \|\tilde{\rho} - \rho\|_\infty + \|\rho - \Sigma\beta^*\|_\infty + \|D\beta^*\|_\infty$$

Using union bounds on the second equation of (5.10) we see that for $\lambda \leq 6\epsilon_0$, we have $P(\|\tilde{\rho} - \rho\|_\infty > \lambda/6) \leq pC\exp\left(-ncs^{-2}\lambda^2\zeta^{-1}\right)$. As $\|D\beta^*\|_\infty \leq sB\|D\|_{\max}$, Lemma 3 alongwith the first equation of (5.10) implies that for $\lambda \leq 12B\epsilon_0$, $P(sB\|D\|_{\max} > \lambda/6) \leq p^2C\exp\left(-ncs^{-2}\lambda^2\zeta^{-1}B^{-2}\right)$. The third component $\rho - \Sigma\beta^* = \frac{1}{n}X'w$ is a linear combination of independent sub-Gaussian errors w. As the columns of X are normalized, invoking property 5.38, we have $P(\|\rho - \Sigma\beta^*\|_\infty > \lambda/6) \leq pC\exp\left(-nc\lambda^2\sigma^{-2}\right)$. Redefining $\zeta = max(\zeta, B^2\zeta, \sigma^2)$ we have

$$\|\tilde{\rho} - \widetilde{\Sigma}\beta^*\|_\infty < \lambda/2 \text{ on } \mathcal{F} \text{ where } P(\mathcal{F}) \geq 1 - p^2C\exp\left(-ncs^{-2}\lambda^2\zeta^{-1}\right)$$

For the remainder of the proof we restrict ourselves to $\mathcal{F}$ adjusting for the probability of $\mathcal{F}^c$. Returning to Equation (5.22), we now have on $\mathcal{F}$,

$$\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}\|_1 \leq \frac{\lambda}{2}||\hat{v}\|_1 + \lambda\|\beta^*\|_1$$

Since $\beta^*_{S^c} = 0$, we know that $\hat{v}_{S^c} = \hat{\beta}_{S^c}$, $\|\beta^*\|_1 = \|\beta^*_S\|_1$. Also for any vector x, we can write $\|x\|_1 = \|x_S\|_1 + \|x_{S^c}\|_1$. Combining these, we have:

$$\frac{1}{2}\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{\beta}_S\|_1 + \lambda\|\hat{v}_{S^c}\|_1 \leq \frac{\lambda}{2}||\hat{v}_S\|_1 + \frac{\lambda}{2}||\hat{v}_{S^c}\|_1 + \lambda\|\beta^*_S\|_1$$

Using the fact that $\|\hat{\beta}_S\|_1 \geq \|\beta^*_S\|_1 - \|\hat{v}_S\|_1$, we now have

$$\hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{v}_{S^c}\|_1 \leq 3\lambda||\hat{v}_S\|_1 \tag{5.23}$$

As $\hat{v}'\widetilde{\Sigma}\hat{v} \geq 0$, we have that on $\mathcal{F}$, $\|\hat{v}_{S^c}\|_1 \leq 3||\hat{v}_S\|_1$. The Restricted Eigenvalue Condition (5.11) immediately implies that on $\mathcal{F}$, $\hat{v}'\Sigma\hat{v} \geq \Omega\|\hat{v}\|_2^2$. Now

$$\begin{aligned}
\hat{v}'\Sigma\hat{v} + \lambda\|\hat{v}\|_1 &= \hat{v}'\widetilde{\Sigma}\hat{v} + \lambda\|\hat{v}_S\|_1 + \lambda\|\hat{v}_{S^c}\|_1 + \hat{v}'D\hat{v} \\
&\leq 4\lambda\|\hat{v}_S\|_1 + \hat{v}'D\hat{v} \text{ using Eqn. (5.23)} \\
&\leq 4\lambda\sqrt{s}\|\hat{v}_S\|_2 + \hat{v}'D\hat{v} \leq 4\lambda\sqrt{s}\|\hat{v}\|_2 + \hat{v}'D\hat{v} \\
&\leq 4\lambda\sqrt{s}\sqrt{\frac{\hat{v}'\Sigma\hat{v}}{\Omega}} + \hat{v}'D\hat{v} \text{ using condition (5.11)} \\
&\leq \frac{\hat{v}'\Sigma\hat{v}}{4} + \frac{16\lambda^2 s}{\Omega} + |\hat{v}'D\hat{v}| \text{ using } 4ab \leq a^2/4 + 16b^2
\end{aligned}$$

The last term on the right hand side is bounded as follows,

$$\begin{aligned}
|\hat{v}'D\hat{v}| &\leq \|D\|_{\max}\|\hat{v}\|_1^2 = \|D\|_{\max}(\|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1)^2 \leq 16||D\|_{\max}\|\hat{v}_S\|_1^2 \text{ on } \mathcal{F} \\
&\leq 16s||D\|_{\max}\|\hat{v}_S\|_2^2 \leq 16s||D\|_{\max}\|\hat{v}\|_2^2
\end{aligned}$$

Using Lemma 3 and the closeness condition (5.10), for $\epsilon \leq min(\epsilon_0, \Omega/64s)$,

$$P(16s\|D\|_{\max} > \Omega/4) = P(\|D\|_{\max} > \Omega/64s) \leq p^2C\exp\left(-nc\epsilon^2\zeta^{-1}\right)$$

With probability at least $1 - p^2C\exp\left(-nc\epsilon^2\zeta^{-1}\right) - p^2C\exp\left(-ncs^{-2}\lambda^2\zeta^{-1}\right)$ we now have

$$\hat{v}'\Sigma\hat{v} + \lambda\|\hat{v}\|_1 \leq \frac{\hat{v}'\Sigma\hat{v}}{4} + \frac{16\lambda^2 s}{\Omega} + \frac{\Omega}{4}\|\hat{v}\|_2^2$$

One more application of the restricted eigenvalue condition (5.11) now yields

$$\frac{\Omega}{2}\|\hat{v}\|_2^2 + \lambda\|\hat{v}\|_1 \leq \frac{16\lambda^2 s}{\Omega}$$

which proves the bounds for both the ℓ_1 and ℓ_2 errors in Theorem 1 $\square$

5.8.2 Proof of Theorem 2

The proof for the sign consistency result of the CoCoLasso is involved. We first present a series of results required to prove Theorem 2.

Lemma 4. *Let $\partial||x||_1$ denotes the sub-gradient of $||x||_1$ for any vector x. Then we have the following results: (a) $\hat{\beta}$ is the optimal solution to $\tilde{f}(\beta) = (1/2)\beta'\widetilde{\Sigma}\beta - \tilde{\rho}'\beta + \lambda|\beta\|_1$ iff there exists a vector $\tilde{u}$ in $\partial\|\hat{\beta}\|_1$ such that*

$$\widetilde{\Sigma}\hat{\beta} - \tilde{\rho} + \lambda\tilde{u} = 0 \tag{5.24}$$

(b) If $|\tilde{u}_j| < 1\ \forall j \in S^c$, then any other optimal solution $\tilde{\beta}$ will have support $S(\tilde{\beta}) \subseteq S$ (c) If we assume that $\widetilde{\Sigma}_{S(\hat{\beta}),S(\hat{\beta})}$ is invertible then under the conditions of part (b), $\tilde{f}(\beta)$ has unique minima

Proof. This lemma is a modified version of (Wainwright, 2009, Lemma 1). We omit the proof as it is exactly analogous to that in the paper. □

Note that the invertibility assumption of part (c) of Lemma 4 needs to hold to establish the uniqueness of the Lasso solution. We now show that this occurs with probability tending to 1. For notational convenience, we define:

$$\delta(\epsilon, \zeta) = p^2 C \exp\left(-cns^{-2}\epsilon^2\zeta^{-1}\right) \tag{5.25}$$

Lemma 5. *$Pr(\widetilde{\Sigma}_{S,S} > 0) \geq 1 - \delta(\epsilon, \zeta)$ for all $\epsilon \leq min(\epsilon_0, C_{\min}/2)$*

Proof. From Equation (5.20), we have

$$\begin{aligned}\Lambda_{\min}(\widetilde{\Sigma}_{S,S}) \geq& \Lambda_{\min}(\Sigma_{S,S}) - |\Lambda_{\max}(-D_{S,S})| \geq C_{\min} - ||D_{S,S}||_2 \\ \geq& C_{\min} - s||D_{S,S}||_{\max} \geq C_{\min} - s||D||_{\max} \geq C_{\min}/2\end{aligned}$$

where the last inequality occurs with probability at least $1-\delta(\epsilon, \zeta)$ for $\epsilon \leq min(\epsilon_0, C_{\min}/2)$ □

Lemma 6. *If $\widehat{\Sigma}$ and $\tilde{\rho}$ satisfy (5.10), then there exists positive constants C, c such that for every $\epsilon \leq min(\epsilon_0, 1/\phi)$,*

$$\begin{aligned}Pr(||F||_\infty \geq \epsilon\phi^2(1-\phi\epsilon)^{-1}) \leq \delta(\epsilon, \zeta) \\ Pr(||H||_\infty \geq \epsilon\phi(2-\gamma)(1-\phi\epsilon)^{-1}) \leq \delta(\epsilon, \zeta)\end{aligned} \tag{5.26}$$

Proof. Let $\eta_1 = ||D_{S,S}||_\infty$ and $\eta_2 = ||D_{S^c,S}||_\infty$. Now, $\sum_{j=1}^{s} |D_{ij}| \leq s||D||_{\max}$ for $(i = 1, \ldots, s)$. Consequently, if $||D||_{\max} \leq \epsilon/s$ then both η_1 and η_2 are less than ϵ. From (5.10) and (5.21), $Pr(\eta_1 \leq \epsilon, \eta_2 \leq \epsilon) \geq 1 - \delta(\epsilon, \zeta)$ for $\epsilon \leq \epsilon_0$. The remainder of the proof follows from (Mai et al., 2012, Lemma A2). □

Proof of Theorem 2 Part (a). We use a Primal Dual Witness construction technique similar to Wainwright (2009) to prove Theorem 2. Let $\hat{\beta}_S$ be the solution to the restricted modified Lasso program i.e.

$$\hat{\beta}_S = \underset{\beta_S}{arg\,min}\ \tilde{f}_S(\beta_S) \text{ where } \tilde{f}_S(\beta_S) = \frac{1}{2}\beta_S' \widetilde{\Sigma}_{S,S}\beta_S - \tilde{\rho}_S'\beta_S + \lambda\|\beta_S\|_1 \tag{5.27}$$

Let $\hat{\beta} = (\hat{\beta}_S', 0'_{(p-s)\times 1})'$ and $\tilde{u} = (\tilde{u}_S', \tilde{u}_{S^c}')'$ where $\tilde{u}_S \in \partial(||\hat{\beta}_S||_1)$ and $\tilde{u}_{S^c}$ is some unspecified $(p-s)\times 1$ vector. From part (a) of Lemma 4, we observe that $\hat{\beta}$ is an optimal solution to (5.7) iff $\{\hat{\beta}, \tilde{u}\}$ satisfies:

$$\begin{aligned} \widetilde{\Sigma}_{S,S}\hat{\beta}_S - \tilde{\rho}_S + \lambda\tilde{u}_S = 0 \\ \widetilde{\Sigma}_{S^c,S}\hat{\beta}_S - \tilde{\rho}_{S^c} + \lambda\tilde{u}_{S^c} = 0 \end{aligned} \tag{5.28}$$

Solving for $\hat{\beta}_S$ and $\tilde{u}_{S^c}$ from Equation (5.28) we have:

$$\hat{\beta}_S = \widetilde{\Sigma}_{S,S}^{-1}(\tilde{\rho}_S - \lambda\tilde{u}_S), \qquad \tilde{u}_{S^c} = \widetilde{G}\tilde{u}_S + \tfrac{1}{\lambda}\left(\tilde{\rho}_{S^c} - \widetilde{G}\tilde{\rho}_S\right) \tag{5.29}$$

From parts (b) and (c) of Lemma 4, we see that $\hat{\beta}$ will be the unique solution to (5.7) if $\widetilde{\Sigma}_{S,S}$ is non-singular and all the entries of $\tilde{u}_{S^c}$ have absolute values less than 1. Lemma 5 provides lower bounds for $Pr(\widetilde{\Sigma}_{S,S} > 0)$. We now derive the bounds for $Pr(||\tilde{u}_{S^c}||_\infty < 1)$. We expand $\tilde{u}_{S^c}$ as :

$$\begin{aligned} \tilde{u}_{S^c} &= G\tilde{u}_S + H\tilde{u}_S + \frac{1}{\lambda}\left((\tilde{\rho}_{S^c} - \rho_{S^c}) + (\rho_{S^c} - G\rho_S) + G(\rho_S - \tilde{\rho}_S) - H\tilde{\rho}_S\right) \\ &= G\tilde{u}_S + H\left(\tilde{u}_s + \frac{1}{\lambda}(\rho_S - \tilde{\rho}_S) - \frac{1}{\lambda}\rho_S\right) \\ &\quad + \frac{1}{\lambda}\left((\tilde{\rho}_{S^c} - \rho_{S^c}) + (\rho_{S^c} - G\rho_S) + G(\rho_S - \tilde{\rho}_S)\right) \end{aligned}$$

Taking the absolute values and using triangular inequalities, we have:

$$\begin{aligned} ||\tilde{u}_{S^c}||_\infty \leq ||G\tilde{u}_S||_\infty + ||H||_\infty \left(1 + \frac{1}{\lambda}||\tilde{\rho}_S - \rho_S||_\infty + \frac{1}{\lambda}||\rho_S||_\infty\right) \\ \frac{1}{\lambda}||\rho_{S^c} - G\rho_S||_\infty + \left(\frac{1}{\lambda}||\tilde{\rho}_{S^c} - \rho_{S^c}||_\infty + \frac{1}{\lambda}||G(\tilde{\rho}_S - \rho_S)||_\infty\right) \end{aligned}$$

We bound each of the four terms on the right hand side separately. The irrepresentable condition (5.13) implies that $||G\tilde{u}_S||_\infty < (1-\gamma)$. It also implies that for $\lambda \leq 4\epsilon_0/\gamma$ we have:

$$\begin{aligned} Pr(&\frac{1}{\lambda}||\tilde{\rho}_{S^c} - \rho_{S^c}||_\infty + \frac{1}{\lambda}||G(\tilde{\rho}_S - \rho_S)||_\infty < \gamma/2) \\ &\geq Pr\left(\frac{1}{\lambda}||\tilde{\rho} - \rho||_\infty < \gamma/4\right) \geq 1 - \delta(\lambda\gamma, \zeta) \end{aligned}$$

where the last inequality follows from taking union bounds on the second equation in (5.10).

The term $(\rho_{S^c} - G\rho_S) = \frac{1}{n}X'_{S^c}(I - X_S(X'_SX_S)^{-1}X'_S)w$ is a linear combination of sub-Gaussian random variables. A direct application of (5.38) yields that $Pr((1/\lambda)\|\rho_{S^c} - G\rho_S\|_\infty \geq \gamma/4) \leq \delta(\lambda\gamma, \zeta)$ where ζ is redefined as maximum of the previous ζ and σ^2.

Without loss of generality, we assume that $\epsilon_0 \leq 1$. Then with probability greater than $1 - \delta(\epsilon, \zeta)$, we can write $||\tilde{\rho}_S - \rho_S||_\infty + ||\rho_S||_\infty \leq ||\tilde{\rho}_S - \rho_S||_\infty + ||\frac{1}{n}X'_Sw||_\infty + ||\frac{1}{n}X'_SX_S\beta^*_S||_\infty \leq 2 + B\psi$ for $\epsilon \leq min(1, \epsilon_0)$. Combining this with Lemma 6, we have, with probability at least $1 - \delta(\epsilon, \zeta)$

$$||H||_\infty \left(1 + \frac{1}{\lambda}||\tilde{\rho}_S - \rho_S||_\infty + \frac{1}{\lambda}||\rho_S||_\infty\right) \leq (1 + \frac{1}{\lambda}(2 + B\psi))\ \frac{\epsilon\phi(2-\gamma)}{(1-\phi\epsilon)} \leq \frac{\gamma}{8}$$

for $\epsilon \leq \epsilon_0^*$ where $\epsilon_0^* = \min(\epsilon_0, \gamma\lambda\phi^{-1}\left(8(2-\gamma)(\lambda + 2 + B\psi) + \gamma\lambda\right)^{-1})$.

Combining all the probabilities and adjusting for the invertibility probability, for $\lambda \leq 4\epsilon_0/\gamma$ and $\epsilon \leq min(\epsilon_0^*, C_{\min}/2)$, we have $Pr(||\tilde{u}_{S^c}||_\infty \geq 1 - \gamma/8) \leq \delta(\lambda\gamma, \zeta) + \delta(\epsilon, \zeta)$. □

Proof of Theorem 2 Parts (b) and (c). Using the expression of $\hat{\beta}_S$ from Equation (5.29), we expand

$$\begin{aligned} \hat{\beta}_S - \beta^*_S =& \widetilde{\Sigma}^{-1}_{S,S}(\tilde{\rho}_S - \rho_S + \frac{1}{n}X'_SX_S\beta^*_S + \frac{1}{n}X'_Sw - \lambda\tilde{u}_S) - \beta^*_S \\ =& F_{S,S}(\tilde{\rho}_S - \rho_S + \frac{1}{n}X'_SX_S\beta^*_S + \frac{1}{n}X'_Sw) \\ &+ \Sigma^{-1}_{S,S}(\tilde{\rho}_S - \rho_S) + \frac{1}{n}\Sigma^{-1}_{S,S}X'_Sw - \lambda\widetilde{\Sigma}^{-1}_{S,S}\tilde{u}_S \end{aligned}$$

We analyze each of the terms above separately. From the definition of sub-Gaussian vectors in (5.10.2) we observe that $\frac{1}{n}\Sigma^{-1}_{S,S}X'_Sw$ is sub-Gaussian with parameter at most $\sigma^2C_{\min}/n$. This implies that $||\frac{1}{n}\Sigma^{-1}_{S,S}X'_Sw\|_\infty$ is less than $\lambda/\sqrt{C_{\min}}$ with probability at

least $1-\delta(\lambda,\zeta)$. Moreover, as $\widetilde{\Sigma}=\Sigma+F$, from Lemma 6 we have with probability at least $1-\delta(\epsilon,\zeta)$, for $\epsilon \leq min(\epsilon_0,(2\phi)^{-1})$:

$$\|\widetilde{\Sigma}_{S,S}\|_\infty \leq \phi + ||F||_\infty \leq \phi + \phi^2\epsilon(1-\phi\epsilon)^{-1} \leq 2\phi$$

The closeness condition for $\tilde{\rho}$ in Equation (5.10) implies that $||\tilde{\rho}_S - \rho_S\|_\infty \leq \lambda$ with probability at least $1-\delta(\lambda,\zeta)$ for $\lambda \leq \epsilon_0$. Following the proof of part (a), we can also conclude that for $\epsilon \leq \epsilon_0$, we have $||\tilde{\rho}_S-\rho_S\|_\infty + ||\frac{1}{n}X_S'X_S\beta_S^*\|_\infty + ||\frac{1}{n}X_S'w\|_\infty \leq (2+B\psi)$ with probability at least $1-\delta(\epsilon,\zeta)$. Therefore,

$$||F_{S,S}(\tilde{\rho}_S - \rho_S + \frac{1}{n}X_S'X_S\beta_S^* + \frac{1}{n}X_S'w)\|_\infty < (2+B\psi)\frac{\phi^2\epsilon}{1-\phi\epsilon} \leq \lambda\phi$$

with probability $1-\delta(\epsilon,\zeta)$ for $\epsilon \leq \lambda\phi^{-1}(\lambda+2+B\psi)^{-1}$. Combining all the probabilities, we have

$$\begin{aligned}
||\hat{\beta}_S - \beta_S^*\|_\infty \leq & ||F_{S,S}(\tilde{\rho}_S - \rho_S + \frac{1}{n}X_S'X_S\beta_S^* + \frac{1}{n}X_S'w)\|_\infty \\
& + \phi||\tilde{\rho}_S - \rho_S\|_\infty + ||\frac{1}{n}\Sigma_{S,S}^{-1}X_S'w\|_\infty + 2\lambda\phi \\
& \leq \lambda\left(4\phi + \frac{1}{\sqrt{C_{\min}}}\right)
\end{aligned}$$

with probability $1-\delta(\lambda,\zeta)-\delta(\epsilon,\zeta)$ for $\epsilon \leq (\epsilon_0, C_{\min}/2, (2\phi)^{-1}, \lambda\phi^{-1}(\lambda+2+B\psi)^{-1})$ and $\lambda \leq \epsilon_0$.

This proves part (b). If $|\beta_{\min}^*| > \lambda(4\phi + \frac{1}{\sqrt{C_{\min}}})$, then the Lasso estimate is sign consistent proving Part (c). □

5.8.3 Proofs of Lemmas 1 and 2

We assume sub-Gaussian additive or multiplicative measurement errors in Section 5.4. The proofs of Lemmas 1 and 2 mainly rely on the properties of sub-Gaussian random variables and vectors which can be found in Appendix 5.10.

Proof of Lemma 1. Let $\Sigma_A = ((\sigma_{a,ij}))$ and b_j denotes the j^{th} column of any matrix B. Then $\widehat{\Sigma}_{add,jk} - \Sigma_{jk} = \frac{1}{n}a_j'x_k + \frac{1}{n}a_k'x_j + (\frac{1}{n}a_j'a_k - \sigma_{a,jk})$. Since $\frac{1}{n}||x_j||_2^2 = 1$ and the entries of a_j are independent and sub-Gaussian with parameter at most τ^2 for all j, property (5.38) implies that $|(1/n)a_j'x_k|$ and $|(1/n)a_j'x_k|$ are each greater than $\epsilon/3$ with probability less than $C\exp\left(-cn\epsilon^2/\tau^2\right)$. Let $z_i = (a_{ij}, a_{ik})'$. Then z_i's are

independent sub-Gaussian vectors with parameter at most τ^2. The tail probability for $\frac{1}{n}a_j'a_k - \sigma_{a,jk}$ can now be made small using Lemma 5.10.1. Hence $\widehat{\Sigma}_{add}$ satisfies (5.10) with $\zeta = max(\tau^4, \tau^2)$ and $\epsilon_0 = c\tau^2$.

We observe that $\tilde{\rho}_{add,j} - \rho_j = \frac{1}{n}a_j'X_S\beta_S^* + \frac{1}{n}a_j'w$. Consequently $|\tilde{\rho}_{add,j} - \rho_j| \leq B\sum_{i=1}^{s}|\frac{1}{n}a_j'x_i| > \epsilon/2$ with probability at most $C\exp(-n\epsilon^2 s^{-2}\tau^{-2}B^{-2})$. Letting $z_i = (a_{ij}, w_i)$, Lemma 5.10.1 can be applied to obtain the tail bound for $\frac{1}{n}a_j'w$. Hence, $\tilde{\rho}_{add}$ satisfies (5.10) with $\zeta = max(\sigma^4, \tau^4, \tau^2B^2)$ and $\epsilon_0 = c\max(\sigma^2, \tau^2)$. □

Proof of Lemma 2. The proof once again relies on Lemma 5.10.1. Let $\Sigma_M = ((\sigma_{m,jk}))$, then

$$\begin{aligned}\widehat{\Sigma}_{mult,jk} - \Sigma_{jk} =& \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(m_{ij}m_{ik} - \mu_j\mu_k - \sigma_{m,jk}) \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk}) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mu_j\mu_k + \sigma_{m,jk}}(\mu_j(m_{ik} - \mu_k) + \mu_k(m_{ij} - \mu_j))\end{aligned}$$

Using the regularity conditions in Equation 5.15, we have,

$$\begin{aligned}|\widehat{\Sigma}_{mult,jk} - \Sigma_{jk}| \leq& \frac{1}{M_{\min}}|(1/n)\sum_{i=1}^{n}x_{ij}x_{ik}((m_{ij} - \mu_j)(m_{ik} - \mu_k) - \sigma_{m,jk})| \\ &+ \frac{\mu_{\max}}{M_{\min}}|(1/n)\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ik} - \mu_k)| \\ &+ \frac{\mu_{\max}}{M_{\min}}|(1/n)\sum_{i=1}^{n}x_{ij}x_{ik}(m_{ij} - \mu_j)|\end{aligned} \tag{5.30}$$

We denote the three terms on the right hand side of (5.30) by T_1, T_2 and T_3 respectively. Note that, if $v = (v_1, v_2, \ldots, v_n)$ where $v_i = x_{ij}x_{jk}$, then $||v||_\infty \leq X_{\max}^2$ As, the errors are once again sub-Gaussian, using Lemma 5.10.1, we see that for $\zeta = \max(\tau^4X_{\max}^4/M_{\min}^2, \tau^2X_{\max}^2\mu_{\max}^2/M_{\min}^2)$ and $\epsilon \leq c\tau^2X_{\max}^2/M_{\min}$ we have:

$$Pr(T_1 \geq \epsilon) \leq C\exp\left(-cn\epsilon^2\zeta^{-1}\right) \text{ for}$$

The terms T_2 and T_3 can be similarly bounded using property (5.38). This proves that $\widehat{\Sigma}_{mult}$ satisfies (5.10). We now show that $\tilde{\rho}_{mult}$ also satisfies (5.10). Recall that

$\tilde{\rho}_{mult,j} - \rho_j = (1/n)(z_j - \mu_j x_j)'y/\mu_j$. As $y = X_S\beta_S^* + w$, we have

$$\begin{aligned}|\tilde{\rho}_{mult,j} - \rho_j| \leq & \frac{1}{\mu_{\min}} \sum_{k=1}^{s} |\frac{1}{n}(z_j - \mu_j x_j)' x_k \beta_k^*| + \frac{1}{\mu_{\min}} |\frac{1}{n}(z_j - \mu_j x_j)' w| \\ \leq & \frac{B}{\mu_{\min}} \sum_{k=1}^{s} |(1/n) \sum_{i=1}^{n} x_{ij} x_{ik} (m_{ij} - \mu_j)| \\ & + \frac{1}{\mu_{\min}} |(1/n) \sum_{i=1}^{n} x_{ij} w_j (m_{ij} - \mu_j)|\end{aligned}$$

Using Lemma 5.10.1, we have for $\zeta = \frac{X_{\max}^2 \max(\tau^2 B^2, \tau^4, \sigma^4)}{\mu_{\min}^2}$ and $\epsilon \leq cX_{max} \frac{\max(\tau^2, \sigma^2)}{\mu_{\min}}$:

$$\begin{aligned}Pr(\frac{1}{\mu_{\min}} |(1/n) \sum_{i=1}^{n} x_{ij} w_j (m_{ij} - \mu_j)| \geq \epsilon/2) \leq C \exp\left(-cn\epsilon^2 \zeta^{-1}\right) \\ Pr(\frac{B}{\mu_{\min}} |(1/n) \sum_{i=1}^{n} x_{ij} x_{ik} (m_{ij} - \mu_j)| \geq \epsilon/2s) \leq C \exp\left(-cn\epsilon^2 s^{-2} \zeta^{-1}\right)\end{aligned}$$

where the last inequality follows from property (5.38). □

5.9 Algorithm for finding the Nearest positive semi-definite matrix

We use an alternating direction method of multipliers to solve for

$$\hat{A} = \underset{A \geq \epsilon I}{\arg\min} ||A - \widehat{\Sigma}||_{\max} \tag{5.31}$$

for any $\epsilon > 0$. We introduce an additional variable B and an equality constraint $B = A - \widehat{\Sigma}$ to rewrite the optimization problem in (5.31) as

$$(\hat{A}, \hat{B}) = \underset{A \geq \epsilon I,\ B = A - \widehat{\Sigma}}{\arg\min} ||B||_{\max} \tag{5.32}$$

To solve (5.32) we will minimize the augmented Lagrangian function:

$$f(A, B, \Lambda) = \frac{1}{2}||B||_{\max} - \langle \Lambda, A - B - \widehat{\Sigma} \rangle + \frac{1}{2\mu}||A - B - \widehat{\Sigma}||_F^2 \tag{5.33}$$

where μ is some penalty parameter, Λ is the Lagrangian matrix and $\langle \cdot, \cdot \rangle$ denotes the matrix inner product which induces the Frobenius norm $||\cdot||_F$. We solve for the minimizer

of $f(A, B, \Lambda)$ iteratively using the following three steps at the i^{th} iteration:

$$\begin{aligned} A \text{ step: } & A_{i+1} = \arg\min_{A \geq \epsilon I} f(A, B_i, \Lambda_i) \\ B \text{ step: } & B_{i+1} = \arg\min_{B} f(A_{i+1}, B, \Lambda_i) \\ \Lambda \text{ step: } & \Lambda_{i+1} = \Lambda_i - \frac{A_{i+1} - B_{i+1} - \widehat{\Sigma}}{\mu} \end{aligned} \tag{5.34}$$

We now provide the closed-form solutions for the first two steps in Equation (5.34). The A step can be simplified as:

$$\begin{aligned} \arg\min_{A \geq \epsilon I} f(A, B_i, \Lambda_i) &= \arg\min_{A \geq \epsilon I} 1/(2\mu)||A - B_i - \widehat{\Sigma}||_F^2 - \langle \Lambda_i, A \rangle \\ &= \arg\min_{A \geq \epsilon I} ||A - B_i - \widehat{\Sigma} - \mu\Lambda_i||_F^2 \end{aligned}$$

The unconstrained solution for the A-step is $B_i + \widehat{\Sigma} + \mu\Lambda_i$. Let for any symmetric matrix Z, Z_ϵ denote the projection of Z into the space of matrices with eigen values greater than ϵ. If $Z = \sum_j \lambda_j p_j p_j'$ denote the spectral decomposition of Z, then we have $Z_\epsilon = \sum_j max(\lambda_j, \epsilon) p_j p_j'$. Hence, the solution for the A-step is given by,

$$A_{i+1} = (B_i + \widehat{\Sigma} + \mu\Lambda_i)_\epsilon \tag{5.35}$$

The B-step is equivalent to:

$$\begin{aligned} &\arg\min_{B} \frac{1}{2}||B||_{\max} + \frac{1}{2\mu}||B - (A_{i+1} - \widehat{\Sigma})||_F^2 - \langle -\Lambda_i, B \rangle \\ &= \arg\min_{B} ||B - (A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i)||_F^2 + \mu||B||_{\max} \end{aligned} \tag{5.36}$$

Let for any symmetric matrix M, $vec_L(M)$ denote the vector containing the lower half elements (including the diagonal) of M. Since, vec_l is an injective mapping, we can define an inverse mapping $mat_l(x)$ such that $mat_l(vec_l(M)) = M$ for any symmetric matrix M. The solution to the B-step is given by

$$B_{i+1} = mat_l(vec_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i) - \ell_1(vec_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i, \mu)))$$

where for any vector x and $\mu > 0$, $\ell_1(x, \mu)$ is the projection of x into the ℓ_1 ball of radius μ. The algorithm to calculate $\ell_1(x, \mu)$ is provided in Duchi et al. (2008).

Algorithm 2 ADMM algorithm for finding the nearest positive semi-definite matrix

1: Input μ and the initial values B_0 and Λ_0

2: At the i^{th} step update:

2.1: (Step A) $A_{i+1} = (B_i + \widehat{\Sigma} + \mu\Lambda_i)_\epsilon$

2.2: (Step B) $B_{i+1} = mat_l(vec_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i) - \ell_1(vec_l(A_{i+1} - \widehat{\Sigma} - \mu\Lambda_i, \mu)))$

2.3: (Step Λ) $\Lambda_{i+1} = \Lambda_i - \frac{A_{i+1} - B_{i+1} - \widehat{\Sigma}}{\mu}$

3: Repeat Step 2 till convergence

5.10 Sub-Gaussian Random Variables

In our analysis of the CoCoLasso estimate, we have assumed that the errors w are independent and identically distributed sub-Gaussian random variables with parameter τ^2. In this Section, we summarize some useful definitions and properties of sub-Gaussian random variables.

Definition 5.10.1. *(Sub-Gaussian random variables, Vershynin 2011) A random variable Z is sub-Gaussian if there exists a finite $\kappa > 0$ such that $\kappa = \sup_{p \geq 1} p^{-1/2}(E|X|^p)^{\frac{1}{p}}$. κ is referred to as the sub-Gaussian norm of Z denoted by $||Z||_\phi$*

Equivalently, a sub-Gaussian random variable Z satisfies

$$P(|Z| > t) \leq 2\exp(-t^2/2\tau^2) \text{ for all } t > 0. \tag{5.37}$$

To avoid ambiguity, we refer to the sub-Gaussian parameter of Z as the smallest τ^2 satisfying (5.37). Following (Vershynin, 2011, Lemma 5.5) we observe that there exists universal constants m and M such that $m||Z||_\phi^2 \leq \tau^2 \leq M||Z||_\phi^2$. We note that if $w = (w_1, w_2, \ldots, w_n)'$ that w_i's are independent zero-centered sub-Gaussian random variables, then weighted sums of w_i are also sub-Gaussian and satisfy an useful property (Vershynin, 2011, Lemma 5.9):

$$||v'w||_\phi^2 \leq K||v||_2^2 \max_i(||w_i||_\phi^2) \tag{5.38}$$

where K is an absolute constant. The tail-probability characterization in (5.37) enables defining sub-Gaussian random vectors in the following sense:

Definition 5.10.2. *(Sub-Gaussian random vectors, Cai et al. 2010 Cai et al. (2010)) A random vector w is said to be sub-Gaussian if there exists $\tau > 0$ such that $Pr(|v'(w - E(w))| > t) \leq 2\exp(-\frac{t^2}{2\tau^2})$ for all $t > 0$ and $||v||_2 = 1$.*

From property (5.10.2) we see that if $w = (w_1, w_2, \ldots, w_n)'$ is a sub-Gaussian vector with parameter τ^2, then each w_i is also sub-Gaussian with parameter at most τ^2. Conversely, if w_i's are independent and sub-Gaussian with parameter τ_i^2, then $w = (w_1, w_2, \ldots, w_n)$ is a sub-Gaussian vector with parameter at most $\tau^2 \leq (KM/m)(\max \tau_i^2)$. We now state and prove an useful result for correlated sub-Gaussian sequences:

Lemma 5.10.1. *Let $z_i = (x_i, y_i)'$ denote independent and identically distributed vectors with zero mean, covariance $\Sigma = ((\sigma_{ij}))$ and sub-Gaussian parameter τ^2. Then there exists absolute constants C and c such that, for every $\epsilon \leq c\tau^2\|a\|_\infty$, we have:*

$$Pr(\frac{1}{n}|\sum_{i=1}^{n} a_i(x_i y_i - \sigma_{12})| \geq \epsilon) \leq C\exp\left(-\frac{nc\epsilon^2}{\tau^4||a||_\infty^2}\right) \tag{5.39}$$

Proof.

$$\begin{aligned}(1/n)\sum_{i=1}^{n} a_i(x_i y_i - \sigma_{12}) &= \frac{1}{4n}\sum_{i=1}^{n} a_i\left((x_i + y_i)^2 - (\sigma_{11} + \sigma_{22} + 2\sigma_{12})\right) \\ &\quad -\frac{1}{4n}\sum_{i=1}^{n} a_i\left((x_i - y_i)^2 - (\sigma_{11} + \sigma_{22} - 2\sigma_{12})\right) \\ &= \frac{1}{2n}\sum_{i=1}^{n} a_i\left((v_1' z_i)^2 - E((v_1' z_1)^2)\right) - \frac{1}{2n}\sum_{i=1}^{n} a_i\left((v_2' z_i)^2 - E((v_2' z_1)^2)\right)\end{aligned}$$

where $v_1 = (1/\sqrt{2}, 1/\sqrt{2})'$ and $v_1 = (1/\sqrt{2}, -1/\sqrt{2})'$. As $||v_k|| = 1$, $v_k' z_1$ is sub-Gaussian with parameter at most τ^2 for $k = 1, 2$. Using the relationship between sub-Gaussian and sub-exponential random variables in (Vershynin, 2011, Lemma 5.14 and Remark 5.18), we see that, for $k = 1, 2$, $(v_k' z_i)^2 - E((v_k' z_1)^2)$ is sub-exponential with parameter at most $c\tau^2$ where c is an absolute constant. As a result $t_i = a_i((v_1' z_i)^2 - E((v_1' z_1)^2)$ is sub-exponential with parameter at most $c\tau^2||a||_\infty$. A direct application of (Vershynin, 2011, Corollary 5.17) now yields for $\epsilon \leq c\tau^2||a||_\infty$,

$$Pr\left(\frac{1}{2n}|\sum_{i=1}^{n} t_i| \geq \epsilon\right) \leq C\exp\left(-\frac{nc\epsilon^2}{\tau^4||a||_\infty^2}\right)$$

□

Chapter 6

Bayesian High Dimensional Changing Linear Regression

6.1 Introduction

Modern statistical modeling and inference continue to evolve and be molded by the emergence of complex datasets, where the dimension of each observation in a dataset substantially exceeds the size of the dataset. Largely due to recent advances in technology, such high dimensional datasets are now ubiquitous in fields as diverse as genetics, economics, neuroscience, public health, imaging, and so on. One important objective of high dimensional data analysis is to segregate a small set of regressors, associated with the response of interest, from the large number of redundant ones. Penalized least square approaches like Lasso (Tibshirani, 1994), SCAD (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), etc. are widely employed for high dimensional regression analysis. Bayesian alternatives typically proceed by using hierarchical priors for the regression coefficients aimed at achieving variable selection. Bayesian variable selection methods include stochastic search variable selection (George and McCulloch, 1993), spike and slab prior (Ishwaran and Rao, 2005), Bayesian Lasso (Park and Casella, 2008), horseshoe prior (Carvalho et al., 2010), shrinkage and diffusion prior (Narisetty and He, 2014) among others.

Most of the aforementioned approaches assume a single underlying model from which the data is generated. Such homogeneity assumptions may be violated in systems where

the variables involved exhibit dynamic behavior and interactions. Common examples include economic time series (Chen and Gupta, 1997; Kezim and Pariseau, 2004; Lenardon and Amirdjanova, 2006), climate change data (Reeves et al., 2007), DNA micro-array data (Baladandayuthapani et al., 2010) and so on. Change point models provide a convenient depiction of such complex relationships by splitting the data based on a threshold variable and using a homogeneous model for each segment. There exists enormous literature on Bayesian methodology addressing various change point problems (see for example Carlin et al., 1992; Barry and Hartigan, 1993; McCulloch and Tsay, 1993; Adams and MacKay, 2007; Turner et al., 2009, among others).

Changing linear regression models are a subclass of change point problems, where the linear model relating the response to the predictors varies over different segments of the data. Segmentation of the dataset is typically based on unknown change points of a threshold variable like time or age or some other contextual variable observed along with the data. Economic datasets constitute a major domain of application of changing linear models. Many economic time series datasets may be collected over different political and financial regimes, thereby containing several change points with respect to the association with the predictors. In a low dimensional setting, Carlin et al. (1992) used Gibbs' sampling techniques for changing linear models to deliver fully Bayesian inference about the location of the change points and the regression coefficients for each segment. When the set of possible predictors is high dimensional, an additional objective is to identify the (possibly different) sparse supports for each segment. Despite the abundance of Bayesian literature on high dimensional regression and on change point models, it appears there is no extant Bayesian work on high dimensional changing linear regression.

This manuscript intends to bridge this gap by proposing a hierarchical method for high dimensional changing linear models. We embed Bayesian variable selection techniques in a change point setup to simultaneously detect the locations of the change points as well as to identify the true sparse support for each of the linear models. We use the newly proposed shrinkage and diffusion priors (Narisetty and He, 2014) for variable selection in a regression framework within each segment. We provide an efficient Gibbs sampler that delivers full posterior inference on the change points, posterior selection probabilities for each variable for all segments, and posterior predictive distributions for

the response. Our fully Bayesian approach is flexible to the choice of variable selection priors and offers the scope for several structural modifications tailored to specific data applications. For example, constraints like grouping the selection of a variable across all the segments can be easily achieved using group selection priors. Other constraints like partial selection within or between the segments can also be accommodated in our setup. We also discuss extensions of our model to identify the number of change points. Numerical studies reveal that for a wide range of scenarios, our proposed method can accurately detect the change points and select the correct set of predictors. We demonstrate the applicability of our method for a macro-economic analysis of Minnesota house price index data. The results strongly favor our change point model over a homogeneous high dimensional regression model.

Classical penalized least square approaches mentioned earlier can also be used in a change point setup. By treating the unknown change points as additional tuning parameters, one can split the data using fixed values of these change points and use some penalized loss function to achieve variable selection for each segment. For example, Lee et al. (to appear) uses the Lasso penalty to estimate the coefficients for each segment. Subsequent application of cross validation or model selection technique yields the optimal change points from a grid of possible values. However, our fully Bayesian approach has several advantages over this. Firstly, the grid search approach is computationally highly inefficient especially for more than one change point. On the other hand, a prior specification for the change points in our Bayesian model enables standard MCMC techniques to efficiently generate posterior samples. Moreover, in many real applications, change in association between variables can occur over a range of the threshold variable. Point estimates of change points obtained from classical approaches fail to accurately depict such scenarios. Bayesian credible intervals obtained from the posterior distributions provide a much more realistic quantification of the uncertainty associated with the location of the change points. This is very difficult to accomplish if one uses the grid search approach.

The rest of the chapter is organized as follows. In Section 6.2 we present our method in details including extensions to unknown number of change points and alternate prior choices. Results from several simulated numerical studies are provided in Section 6.3. In Section 6.4 we present the details of a house price index data analysis using our change

point method. We conclude in Section 6.5 with a brief review and pointers to future research.

6.2 Method

We consider a traditional high dimensional setup with the $n \times 1$ response vector $y = (y_1, y_2, \ldots, y_n)'$ and corresponding $n \times p$ covariate matrix $X = (x_1, x_2, \ldots, x_n)'$ where p can be larger than n. We further assume that for every observation y_i we observe another quantitative variable t_i such that the association between y_i and x_i depends on the values of t_i. In a linear regression setup, this dynamic relationship between the response y_i and the corresponding $p \times 1$ vector of covariates x_i can be expressed as $E(y_i \,|\, x_i, t_i) = x_i'\beta_k$ for all i such that $\tau_{k-1} < t_i < \tau_k$ where $\tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = n$. The change-points $\tau_1, \tau_2, \ldots, \tau_K$ are typically unknown while the number of change-points K may or may not be known depending on the application.

As the number of regressors (p) is large, our goal is to select the relevant variables for this regression. However, for this changing linear regression, the set of relevant regressors may depend on the value of the threshold variable t and variable selection procedures applied disregarding the dependence on t can lead to erroneous variable selection. Let S_k denote the support of β_k where $s_k = |S_k|$ is typically much less than p. Our goal is to simultaneously detect the change-points τ_k and estimate S_k for all $k = 1, 2, \ldots, K$. We initially assume only one change-point τ, i.e., $K = 1$. Extensions to more than one (and possibly an unknown number of) change points are discussed later in Section 6.2.2.

6.2.1 One Change Point Model

We assume a changing linear regression model

$$y_i = \begin{cases} x_i'\beta_1 + \epsilon_i \text{ if } t_i \leq \tau \\ x_i'\beta_2 + \epsilon_i \text{ if } t_i > \tau \end{cases} \tag{6.1}$$

where β_1, β_2 are both sparse $p \times 1$ vectors such that $\beta_1 \neq \beta_2$ and $\epsilon_i \sim N(0, \sigma^2)$ denotes the independent and identically distributed noise. To accomplish variable selection both before and after the change point, we use Bayesian shrinking and diffusion (BASAD) priors proposed in Narisetty and He (2014) for β_1 and β_2. To be

specific, we assume $\beta_k \,|\, Z_k, \sigma^2 \sim N(0, \sigma^2 diag(\gamma_{1k} Z_k + \gamma_{0k}(1 - Z_k)))$ for $k = 1, 2$ where $Z_k = (Z_{k1}, Z_{k2}, \ldots, Z_{kp})'$ is a $p \times 1$ vector of zeros and ones. The hyper-parameters γ_{0k} and γ_{1k} are scalars chosen to be very small and very large respectively. Hence, β_{kj}—the j^{th} component of β_k—is assigned a shrinking (concentrated around zero) prior if Z_{kj} equals 0 and a diffusion (flat) prior if $Z_{kj} = 1$. The Z_{kj}'s are assumed to be *a priori* independent, each following $Bernoulli(q_k)$. Hence q_k controls the prior model size for the k^{th} segment. The choices for the hyper-parameters γ_{0k}, γ_{1k} and q_k are discussed in Section 6.3. We assume a uniform prior for the change-point τ and a conjugate Inverse Gamma prior for the noise variance σ^2. The full Bayesian model can now be written as:

$$\prod_{i:t_i \leq \tau} N(y_i \,|\, x_i' \beta_1, \sigma^2) \prod_{i:t_i > \tau} N(y_i \,|\, x_i' \beta_2, \sigma^2) \times Unif(\tau \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times$$
$$\prod_{k=1}^{2} \left(N(\beta_k \,|\, 0, \sigma^2 diag(\gamma_{1k} Z_k + \gamma_{0k}(1 - Z_k))) \times \prod_{j=1}^{p} Bernoulli(Z_{kj} \,|\, q_k) \right) \quad (6.2)$$

We use Gibbs' sampler to obtain posterior samples of τ, σ^2, β_k and Z_k for $k = 1, 2$. Let $\tau \,|\, \cdot$ denote the full conditional distribution of τ in the Gibbs sampler. We use similar notation to denote the other full conditionals. Let $U_1 = \{i \,|\, t_i \leq \tau\}$ and $U_2 = \{i \,|\, t_i > \tau\}$. For $k = 1, 2$, let Y_k and X_k denote the response vector and covariate matrix obtained by stacking up the observations U_k. From the full joint distribution in (6.2), we have

$$\beta_k \,|\, \cdot \sim N(V_k X_k' Y_k,\ \sigma^2 V_k) \text{ where } V_k = (X_k' X_k + diag(\gamma_{1k} Z_k + \gamma_{0k}(1 - Z_k))^{-1})^{-1}$$

$$\sigma^2 \,|\, \cdot \sim IG(a_\sigma + n/2, b_\sigma + + \frac{1}{2} \sum_{k=1}^{2} ||Y_k - X_k \beta_k||^2)$$

$$p(\tau \,|\, \cdot) \propto \prod_{k=1}^{2} \prod_{i \in U_k} N(y_i \,|\, x_i' \beta_k, \sigma^2) \times Unif(\tau \,|\, a_\sigma, b_\sigma)$$

$$Z_{kj} \,|\, \cdot \sim Bernoulli \left(\frac{q_k \ \phi(\beta_{kj} / \sqrt{\sigma^2 \gamma_{1k}})}{q_k \ \phi(\beta_{kj} / \sqrt{\sigma^2 \gamma_{1k}}) + (1 - q_k) \ \phi(\beta_{kj} / \sqrt{\sigma^2 \gamma_{0k}})} \right)$$

where $\phi(\cdot)$ denotes the density of standard normal distribution. We observe that the full conditionals of β_k, Z_{kj} and σ^2 follow conjugate distributions and are easily updated via the Gibbs sampler. Only $p(\tau \,|\, \cdot)$ does not correspond to any standard likelihood and we use a random walk Metropolis-Hastings step within the Gibbs sampler to update τ.

6.2.2 Multiple change points

So far we have limited our discussion to the presence of only one change point. However, our method can be easily extended to multiple change points. If we have K change points $\tau_1 < \ldots < \tau_K$, the joint likelihood in (6.2) can be generalized to

$$\prod_{k=1}^{K} \left(\prod_{i:\tau_{k-1}<t_i\leq\tau_k} N(y_i \,|\, x_i'\beta_k, \sigma^2) \times N(\beta_k \,|\, 0, \sigma^2 diag(\gamma_{1k}Z_k + \gamma_{0k}(1-Z_k))) \times \prod_{j=1}^{p} Bernoulli(Z_{kj} \,|\, q_k) \right) \times p(\tau_1, \tau_2, \ldots, \tau_K) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \tag{6.3}$$

To ensure identifiability of the change points, the prior $p(\tau_1, \tau_2, \ldots, \tau_K)$ should be supported on $\tau_1 < \tau_2 < \ldots < \tau_K$. To accomplish this we choose $p(\tau_1, \tau_2, \ldots, \tau_K)$ as the density of order statistics of a sample of size K from $Unif(a_\tau, b_\tau)$. The Gibbs sampler remains essentially same as in Section 6.2 with the Metropolis random walk step now being used to update the entire change point vector $(\tau_1, \tau_2, \ldots, \tau_K)'$.

6.2.3 Determining the number of change points

Often in applications, the number of change points is unknown. In our fully Bayesian approach this can potentially be handled by adding a prior for the number of change points (K). Introducing this additional level of hierarchy comes with the caveat that different values of K yields parameter sub-spaces of different sizes and interpretations. To elucidate, a one change point model splits the data into two segments, with separate coefficient vectors β_1 and β_2, creating a parameter space of dimension $2p$ whereas a no change point model has a single β of dimension p with a possible interpretation that it is some average of β_1 and β_2 over the two segments. Therefore, a Markov Chain Monte Carlo sampling for K will involve jumping within and between different sub-spaces.

Green (1995) proposed the extremely general and powerful reversible jump MCMC (RJMCMC) sampler for sampling across multiple parameter spaces of variable dimensions. We can seamlessly adopt an RJMCMC joint sampler to obtain the posterior distribution for the number of change points. When naively implemented, RJMCMC experiences poor acceptance rates for transitions to parameter sub-spaces with different dimensionality. This leads to widely documented convergence issues (Green and Hastie,

2009; Fan and Sisson, 2011). The problem will be exacerbated in our setup due to the high dimensionality of the parameter spaces.

Several improvements and alternatives to RJMCMC have been proposed over the years including efficient proposal strategies to effectuate frequent cross-dimensional jumps (Richardson and Green, 1997; Brooks et al., 2003; Ehlers and Brooks, 2008; Farr et al., 2015), product space search (Carlin and Chib, 1995; Dellaportas et al., 2002) and parallel tempering (Littenberg and Cornish, 2009). All these approaches can be adapted in our setup to determine the number of change points. However, many of these approaches are accompanied by their own computational burden such as running several chains or *a priori* obtaining posterior distributions for each individual model before running the joint sampler. We concur with Han and Carlin (2001) and Hastie and Green (2012) that it is often expedient to use simpler model selection approaches based on individual models. Hence, popular Bayesian model comparison metrics like DIC (Spiegelhalter et al., 2002) and l-measure (Gelfand and Ghosh, 1998) remain relevant for selecting the number of change points. For example, if θ is the complete set of parameters associated with the model, for each K we can compute the DIC score

$$\text{DIC} = 2E\left(D(y \,|\, \theta) \,|\, y\right) - D\left(y \,|\, E(\theta \,|\, y)\right) = E\left(D(y \,|\, \theta) \,|\, y\right) + p_D \tag{6.4}$$

where $D(y \,|\, \theta)$ is the deviance function and $p_D = E\left(D(y \,|\, \theta) \,|\, y\right) - D\left(y \,|\, E(\theta \,|\, y)\right)$ is interpreted as effective model size. Hence, DIC penalizes more complex models and is particularly suitable for our change point context where a higher number of change points will lead to overfitting. Parallel computing can be utilized to simultaneously run the MCMC sampler for different values of K and then the optimal K can be selected as the one yielding the lowest DIC score.

All the methods discussed here for selecting the number of change points discussed here can be used in conjunction with our approach. It is prudent to predicate the choice on the nature of the application at hand and the computational resources available.

6.2.4 Alternate prior choices

We observe from Equation 6.2 that, conditional on the value of the change point τ, the joint likelihood can be decomposed into individual likelihoods for the regression

before and after the change point along with the corresponding priors for the regression coefficients. This allows much flexibility in the choice of priors for the regression coefficients.

We have focused on the BASAD prior. One can also use other priors to achieve variable selection. For example, using Laplace (double exponential) priors for the β_k's will yield a Bayesian Lasso (Park and Casella, 2008) with change point detection. To facilitate the discussion, consider the one change point model. By using the Laplace prior, the full hierarchical specification for the coefficient vectors β_k for $k = 1, 2$ can be specified as:

$$\begin{aligned} \beta_k \,|\, \sigma^2, \eta_k &\overset{ind}{\sim} N(0, \sigma^2\, diag(\eta_k)) \text{ where } \eta_k = (\eta_{k1}, \eta_{k2}, \ldots, \eta_{kp})' \\ \eta_{kj} \,|\, \lambda_k &\overset{ind}{\sim} Exp(\lambda_k^2/2) \text{ and } \lambda_k^2 \sim Gamma(r_k, s_k) \end{aligned} \tag{6.5}$$

The prior specification for σ^2 and τ can be kept same as in (6.2). The Gibbs sampler for the Bayesian Lasso provided in Park and Casella (2008) can now be used to sample from the following full conditionals:

$$\begin{aligned} \beta_k \,|\, \cdot &\sim N(V_k X_k' y_k, \sigma^2 V_k) \text{ where } V_k = (X_k' X_k + diag(\eta_k)^{-1})^{-1} \\ \sigma^2 \,|\, \cdot &\sim IG(a_\sigma + n/2,\, b_\sigma + \frac{1}{2}\sum_{k=1}^{2} ||Y_k - X_k\beta_k||^2) \\ 1/\eta_{kj} \,|\, \cdot &\sim \text{ Inv-Gauss } \left(\sqrt{\frac{\lambda_k^2\sigma^2}{\beta_{kj}^2}}, \lambda_k^2\right) \\ \lambda_k^2 \,|\, \cdot &\sim Gamma(r_k + p/2, s_k + \frac{1}{2}||\eta_k||_2^2) \\ p(\tau \,|\, \cdot) &\propto \prod_{k=1}^{2} \prod_{i \in U_k} N(y_i \,|\, x_i'\beta_k, \sigma^2) \times Unif(\tau \,|\, a_\sigma, b_\sigma) \end{aligned}$$

Additional information regarding grouping or structuring of the variables is often available in the context of variable selection. In the presence of a change point, additional constraints can specify grouped selection both within or between the β_k's. For example, in a single change point setup, it may be plausible that the set of relevant variables remains unchanged before and after the change point, with change occurring only with respect to the strength of association between y_i and x_i. Such additional structural constraints both within and across β_k's can easily be accommodated in our

setup via a suitable choice of prior. To elucidate, we can rewrite (6.1) as $y_i = z_i(\tau)'\zeta + \epsilon_i$ where $z_i = (I(t_i \leq \tau)x_i', I(t_i > \tau)x_i')'$ and $\zeta = (\beta_1', \beta_2')'$. To incorporate the constraint that β_1 and β_2 share the same support, one can use a Bayesian group lasso (Raman et al., 2009) with M-Laplace priors on the groups $\zeta_j = (\beta_{1j}, \beta_{2j})'$ for $j = 1, 2, \ldots, p$. The M-Laplace prior

$$p(\zeta_j \,|\, \sigma^2, \lambda^2) \propto \frac{2\lambda^2}{\sigma^2} \exp(-\sqrt{\frac{2\lambda^2}{\sigma^2}}||\zeta_j||_2)$$

has a convenient two-step hierarchical specification:

$$\zeta_j \,|\, \eta_j \overset{ind}{\sim} N(0, \sigma^2\eta_j I);\ \eta_j \,|\, \lambda^2 \overset{ind}{\sim} Gamma(3/2, \lambda^2);\ \lambda^2 \sim Gamma(r, s) \tag{6.6}$$

The full conditional distributions of the parameters provided in Raman et al. (2009) can now be used to implement the Gibbs sampler with the additional Metropolis random walk step for updating the change point τ. Any other information like hierarchical selection or anti-hierarchical selection both within and between the β_k's can also be accommodated via suitable priors.

Often, in real data applications, prior knowledge dictates the inclusion of certain variables in the model and variable selection is sought only for the remaining variables. Such constraints can be easily achieved in our setup by using standard Gaussian prior for that specified subset and BASAD prior for the remaining variables.

6.2.5 Variable selection after MCMC

When there are finitely many candidate models, Bayesian model selection typically proceeds by selecting the candidate model with the highest posterior probability. However, in our setup the regression coefficients are continuous. For variable selection, we use the median probability model (Barbieri and Berger, 2004) which is computationally easy and is optimal in terms of prediction. To be specific, β_{kj} is included in the model if the posterior probability of $Z_{kj} = 1$ is greater than 0.5.

6.3 Numerical Studies

We conducted numerical experiments to assess the performance of our method both for single and multiple change points. For all the simulation studies we used $100,000$

MCMC iterations. Multiple chains were run with different choices of initial values and convergence was typically achieved within the first $20,000$ iterations. Nevertheless, we discarded the first $50,000$ as burn-in and used the subsequent $50,000$ samples for inference.

6.3.1 One change point

We assume $t_i = i$ and generate data from the model $y_i = N(x_i'\beta_1, \sigma^2)$ for $i \leq \tau$ and $y_i = N(x_i'\beta_2, \sigma^2)$ for $i > \tau$ where $\beta_1 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$ and $\beta_2 = -\beta_1$. The rows of X were independent and identically distributed normal random variables with zero mean and covariance Σ_X. Two structures were used for Σ_X — auto-regressive (AR) with $\Sigma_{X,ij} = 0.5^{|i-j|}$ and compound symmetry (CS) with $\Sigma_{X,ij} = 0.5 + 0.5I(i = j)$. The noise variance σ^2 was fixed at 1 and the sample size was chosen to be 200. Two different model sizes — $p = 250$ and $p = 500$ were used. The change point τ was chosen to vary between 50.5, 100.5 and 150.5. Since the sample size is 200, these three choices of τ respectively correspond to changes in the regression model at the initial, middle or later portion of the data. We used three different prior choices for the coefficients — the BASAD prior, the Bayesian Lasso prior (Park and Casella, 2008) and the Bayesian Group Lasso prior (Raman et al., 2009). The last choice was used to investigate any possible benefits of using a grouped variable selection as it is known that β_1 and β_2 has same support. The range of the uniform prior for τ was chosen to be $(20, 180)$ and a normal proposal density with tuning variance of 0.1 was used for the Metropolis update of τ. The prior for σ^2 was chosen to be $IG(2, 1)$. The hyper-parameters γ_{0k}, γ_{1k} and q_k were chosen as follows. Let τ_0 denote the initial estimate for τ. Then $n_1 = [\tau_0]$ and $n_2 = n - n_1$ denotes the initial sample sizes for the two segments. We used

$$\gamma_{0k} = \frac{\hat{\sigma}_k^2}{10n_k} \qquad , \qquad \gamma_{1k} = \hat{\sigma}_k^2 \ \max\left(\frac{p^{2.1}}{100n_k}, \log n_k\right)$$

where $\hat{\sigma}_k^2$ was the sample variance of Y_k for $k = 1, 2$. The hyper-parameters q_k were chosen such that the prior model sizes $\sum_{j=1}^{p} Z_{kj}$ were greater than $\min(p-1, \max(10, \log n_k))$ with probability 0.1. These choices of γ_{0k}, γ_{1k} and q_k were adapted from Narisetty and He (2014).

The posterior median estimates of the change points for all the scenarios are provided in Tables 6.1 ($p = 250$) and 6.2 ($p = 500$). We observe that all the 3 models estimate

Table 6.1: Single change point model with $n = 200$ and $p = 250$: Posterior median estimates (and 95% confidence intervals) of τ using BASAD, Bayesian Lasso (BL) and Bayesian Group Lasso (BGL) priors.

	True τ	BASAD	BL	BGL
AR	50.5	50.5 (50.0, 51.0)	50.4 (48.9, 51.8)	50.5 (49.5, 51.5)
	100.5	100.5 (100.0, 101.0)	100.5 (100.0, 101.0)	100.5 (100.0, 101.0)
	150.5	150.5 (150.0, 151.0)	150.5 (150.0, 151.0)	150.5 (150.0, 151.0)
CS	50.5	50.1 (49.1, 51.9)	50.0 (49.1, 51.8)	50.1 (49.1, 51.9)
	100.5	100.5 (100.0, 101.0)	100.5 (100.0, 101.0)	100.5 (100.0, 101.0)
	150.5	150.5 (150.0, 151.0)	150.2 (148.6, 151.0)	150.4 (149.1, 151.0)

Table 6.2: Single change point model with $n = 200$ and $p = 500$: Posterior median estimates (and 95% confidence intervals) of τ using BASAD, Bayesian Lasso (BL) and Bayesian Group Lasso (BGL) priors.

	True τ	BASAD	BL	BGL
AR	50.5	50.5 (50.0, 51.0)	49.7 (47.6, 53.7)	50.4 (48.2, 53.1)
	100.5	101.3 (99.3, 102.0)	101.3 (99.1, 103.9)	101.0 (99.1, 103.3)
	150.5	150.5 (150.0, 151.0)	151.3 (148.7, 152.9)	150.7 (150.0, 152.8)
CS	50.5	50.3 (49.1, 51.0)	50.1 (48.5, 51.8)	50.2 (49.1, 51.5)
	100.5	100.3 (99.1, 101.0)	100.0 (99.1, 100.9)	99.9 (99.1, 100.9)
	150.5	150.5 (150.0, 151.0)	150.5 (150.0, 151.0)	150.5 (150.0, 151.0)

the change point with high accuracy.

We then turn our attention to variable selection. Let C_k and IC_k denote the number of true and false regressors respectively selected for the k^{th} segment of the data for $k = 1, 2$. As discussed earlier, we used a cut-off of 0.5 for the posterior probability of the binary Z_{kj}'s in the BASAD model to select the variables. The Bayesian Lasso and Group Lasso are devoid of such binary selection parameters and variable selection was based on the posterior confidence intervals, i.e., β_{kj} was not selected if its posterior confidence interval covered zero. Table 6.3 provides the C_k and IC_k numbers for each method for $p = 250$. We observe that the BASAD prior achieves perfect variable selection for all the scenarios while the Bayesian LASSO often misses out on a true

Table 6.3: Single change point model with $n = 200$ and $p = 250$: Number of correct and incorrect predictors selected by BASAD, Bayesian Lasso (BL) and Bayesian Group Lasso (BGL). Cases where any method missed at least one true regressor are highlighted using *.

			AR			CS		
		True	BASAD	BL	BGL	BASAD	BL	BGL
$\tau = 50.5$	C_1	3	3	3	3	3	2*	3
	C_2	3	3	3	3	3	3	3
	IC_1	0	0	0	0	0	0	0
	IC_2	0	0	2	2	0	2	1
$\tau = 100.5$	C_1	3	3	3	3	3	3	3
	C_2	3	3	3	3	3	3	3
	IC_1	0	0	0	0	0	0	0
	IC_2	0	0	0	0	0	0	0
$\tau = 150.5$	C_1	3	3	3	3	3	3	3
	C_2	3	3	2*	3	3	2*	3
	IC_1	0	0	0	0	0	2	1
	IC_2	0	0	0	0	0	0	0

variable and includes some incorrect predictors. The Bayesian Group LASSO always selects the true set of regressors but often includes one false regressor. For $p = 500$, (Table 6.4), the BASAD once again selects the exact set of predictors. However, the performance of the Bayesian Group LASSO and especially the Bayesian LASSO worsens with the latter often being able to select only one correct predictor.

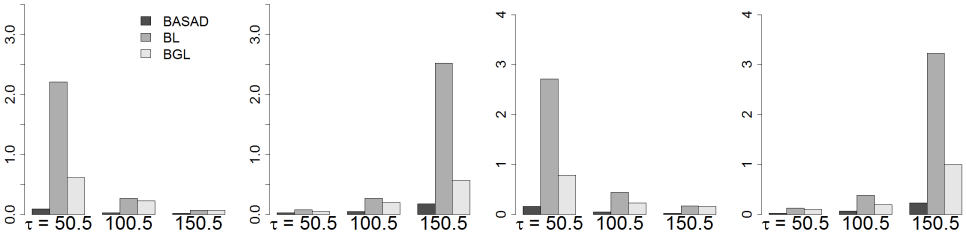
In addition to the variable selection metrics, we also assess the three methods based on the coefficient estimates for the true predictors using the Squared Error truncated on the true support i.e.

$$SE_k = ||\beta_k[\mathcal{S}_k] - \hat{\beta}_k[\mathcal{S}_k]||_2^2 \text{ for } k = 1, 2$$

where $\mathcal{S}_k$ denotes the true support of β_k and $\hat{\beta}_k$ denote its posterior estimate. Figures 6.1 and 6.2 plots the SE_k numbers for $p = 250$ and $p = 500$ respectively. We observe that BASAD stands out with uniformly lowest SE numbers across all scenarios. It is

Table 6.4: Single change point model with $n = 200$ and $p = 500$: Number of correct and incorrect predictors selected by BASAD, Bayesian Lasso (BL) and Bayesian Group Lasso (BGL). Cases where any method missed at least one true regressor are highlighted using *.

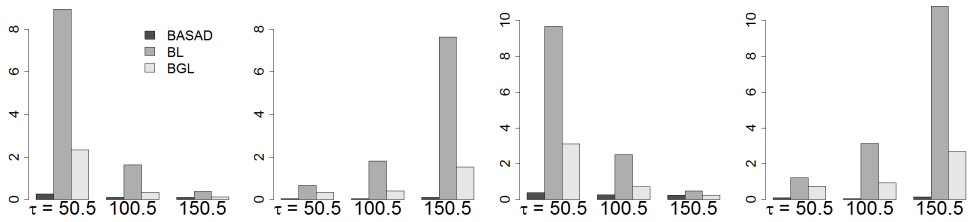
			AR			CS		
		True	BASAD	BL	BGL	BASAD	BL	BGL
$\tau = 50.5$	C_1	3	3	1*	3	3	1*	2*
	C_2	3	3	3	3	3	3	3
	IC_1	0	0	0	0	0	0	0
	IC_2	0	0	0	0	0	0	0
$\tau = 100.5$	C_1	3	3	3	3	3	3	3
	C_2	3	3	3	3	3	3	3
	IC_1	0	0	0	0	0	0	0
	IC_2	0	0	0	0	0	0	0
$\tau = 150.5$	C_1	3	3	3	3	3	3	3
	C_2	3	3	1*	3	3	0*	3
	IC_1	0	0	0	0	0	0	0
	IC_2	0	0	0	0	0	0	0



(a) SE_1: AR (b) SE_2: AR (c) SE_1: CS (d) SE_2: CS

Figure 6.1: Single change point model with $p = 250$: Posterior median estimates of the truncated Squared Error (SE_k) for β_1 and β_2 using BASAD, Lasso and Group Lasso (GL) priors

important to note that, SE_1 tends to be higher when $\tau = 50.5$ while SE_2 is higher when $\tau = 150.5$. This behavior is expected as for $\tau = 50.5$, sample size for estimating β_1 is



(a) SE_1: AR (b) SE_2: AR (c) SE_1: CS (d) SE_2: CS

Figure 6.2: Single change point model with $p = 500$: Posterior median estimates of the truncated Squared Error (SE_k) for β_1 and β_2 using BASAD, Lasso and Group Lasso (GL) priors

effectively 50 while that for β_2 is 150. The Bayesian Lasso seems to be worst impacted by this effective sample size. It performs worse for the Compound Symmetry covariance structure and for higher model size ($p = 500$). The Bayesian Group Lasso, enjoying the additional knowledge of the structural constraint, conceivably performs better than the Bayesian Lasso. However, the BASAD seems to be least impacted by effective sample size, model size or covariance structure and produces accurate variable selection, change point detection and estimation across all scenarios.

6.3.2 Two change points

We demonstrate the applicability of our method to multiple change points using a two change point setup. The three coefficient vectors are given by

$$\beta_1 = (3, 0, 0, \ldots, 0)',\ \beta_2 = (3, 1.5, 0, 0, \ldots, 0)', \beta_3 = (3, 1.5, 0, 0, 2, 0, 0, \ldots, 0)'$$

Three pairs of values for the change points (τ_1, τ_2) are chosen — $(50.5, 100.5)$, $(50.5, 150.5)$ and $(100.5, 150.5)$. Other specifications including sample size, model size and covariance of the predictors are kept unchanged from Section 6.3.1. We observed in Figures 6.1 and 6.2 that the Bayesian Lasso becomes erratic for small effective sample sizes. For two change points, the effective sample size is further lowered and the Bayesian Lasso faced convergence issues in the MCMC sampler. The Bayesian Group Lasso also cannot be used here as the coefficient vectors for different segments do not share a common support. Hence, we only present the results for the BASAD prior.

Table 6.5 presents the change-point estimates for all the scenarios. We observe that all the change points are accurately estimated. Turning to variable selection, once

Table 6.5: Two change point model: Posterior median and 95% confidence intervals of the change points.

			(50.5,100.5)	(50.5,150.5)	(100.5,150.5)
$p = 250$	AR	τ_1	50.8 (49.2, 53.9)	50.5 (48.3, 52)	100.6 (100, 102.7)
		τ_2	101.5 (101, 103.3)	148.6 (146.9, 151.7)	150.1 (149, 152.5)
	CS	τ_1	50.6 (45.7, 52.9)	50.8 (49.1, 52.9)	98.6 (92.6, 100.8)
		τ_2	101.3 (97.3, 104.1)	151.3 (148.1, 152.7)	151 (145.4, 152.8)
$p = 500$	AR	τ_1	50.6 (45.7, 52.9)	50.8 (49.1, 52.9)	98.6 (92.6, 100.8)
		τ_2	101.3 (97.3, 104.1)	151.3 (148.1, 152.7)	151 (145.4, 152.8)
	CS	τ_1	47.2 (43.3, 51.8)	50.4 (49.1, 51.9)	101.5 (97.6, 103.2)
		τ_2	102.1 (97.9, 105.8)	147.4 (142.1, 151.8)	146.2 (140.3, 151.3)

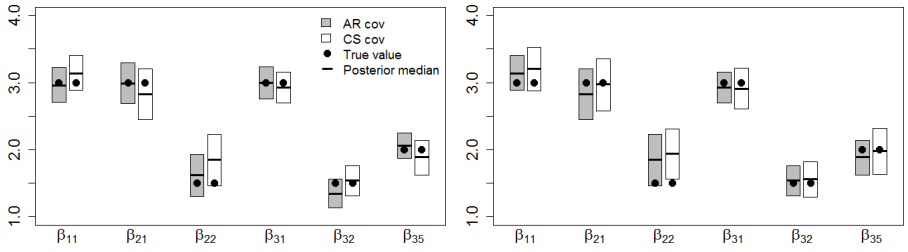
again, BASAD identified the exact set of predictors across all scenarios. Figure 6.3 provides the estimates of the non-zero coefficients in the model. We observe that with the exception of the the second entry of β_2, all the non-zero coefficients were well within the posterior confidence intervals. Overall, we observe that even for multiple change points, our methodology can accurately detect the change points, identify the correct sets of predictors, and estimate the regression coefficients.

6.4 Minnesota House Price Index Data

In this section we apply our method to conduct an empirical analysis of Minnesota house price index data.

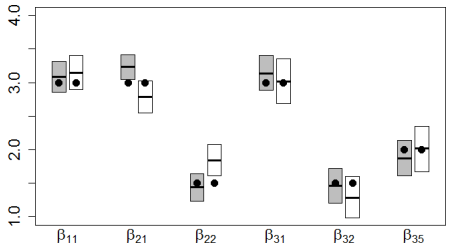
6.4.1 Literature Review

The expansive literature on statistical analysis of house price data can be broadly classified into two major subdivisions based on their objectives. The first category of articles is aimed at forecasting individual house prices based on the constituent characteristics of the houses. Such hedonic regression models usually incorporate information regarding

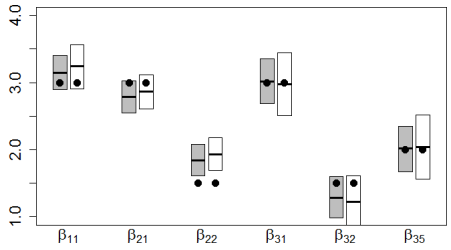


(a) $\tau_1 = 50.5$, $\tau_2 = 100.5$, $p = 250$

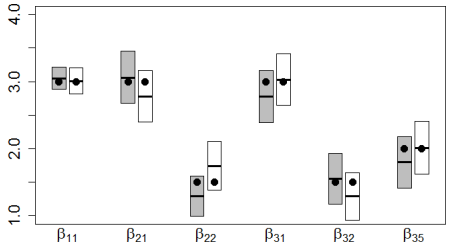
(b) $\tau_1 = 50.5$, $\tau_2 = 100.5$, $p = 500$



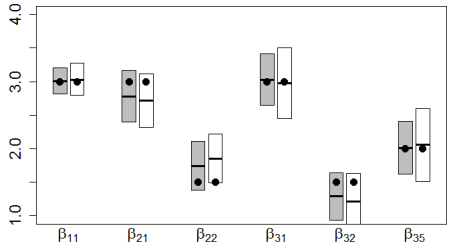
(c) $\tau_1 = 50.5$, $\tau_2 = 150.5$, $p = 250$



(d) $\tau_1 = 50.5$, $\tau_2 = 150.5$, $p = 500$



(e) $\tau_1 = 100.5$, $\tau_2 = 150.5$, $p = 250$



(f) $\tau_1 = 100.5$, $\tau_2 = 150.5$, $p = 500$

Figure 6.3: Two change point model: Posterior median estimates and 95% confidence intervals of the non-zero entries of β_1, β_2 and β_3. β_{kj} denotes the k entry of β_k for $k = 1, 2, 3$

the structure, location, neighborhood, and selling history of the house to determine its price range (see Malpezzi, 2003, for a comprehensive review on hedonic models). The second class of analysis focuses on understanding how real estate prices impact or are impacted by the economy of a country or a region. Until very recently this subdivision received relatively scant attention because interaction between housing market and macroeconomic variables was often deemphasized (Leung, 2004). The US sub-prime mortgage crisis between 2007 and 2009, triggered by the collapse of the housing market has resuscitated interest on studying this relationship.

Several empirical analyses furnish evidence for co-movements of house price indices and other macro-economic variables like Gross Domestic Product (GDP), consumer price indices, unemployment rates, interest rates, stock price indices, and so on (Apergis, 2003; Hofmann, 2003; Tsatsaronis and Zhu, 2004; Otrok and Terrones, 2005; Renigier-Biozor and Winiewski, 2013; Panagiotidis and Printzis, 2015). Multivariate regression models have been used to understand the relationship between these macro-economic variables and house price index (hpi) in Ukraine (Mavrodiy, 2005), Sweden (Strömberg et al., 2011), and Malaysia (Ong and Chang, 2013). These analyses often assume a single underlying time-homogeneous relationship between hpi and the explanatory variables. Such an assumption may be far-fetched in reality, where correlation between hpi and its macroeconomic determinants may exhibit differential trends over time. For example, as noted in Ahamada and Diaz Sanchez (2013), the US stock market crash in the 'Internet bubble burst' of 2001-2002 was not accompanied by plummeting house prices whereas in the sub-prime mortgage crisis in 2007-2009, stocks and house prices witnessed simultaneous collapse.

While the impact of macroeconomic variables on US house prices has been analysed in the literature (Case et al., 2001; Catte et al., 2004) any relevant literature focusing on similar analysis at state level has eluded us. As housing markets are local in nature (Garmaise and Moskowitz, 2002), a state level macro-analysis may reveal trends not reflected in a similar nationwide study. The state of Minnesota is home to 18 Fortune 500 companies (`http://mn.gov/deed/business/locating-minnesota/companies-employers/fortune500.jsp`) and has the second highest number of Fortune 500 companies per capita. Furthermore, the Minneapolis-St. Paul metropolitan area hosts the highest number of Fortune 500 companies per capita among the 30 largest

metropolitan areas in US. Hence, local industries may play a significant role in determining real estate prices in Minnesota. We use a multivariate regression model with change points to investigate the relationship between the state-level hpi of Minnesota and both local and national macro-economic variables.

6.4.2 Data and Model

For our analysis, we use quarterly Minnesota hpi data published by the Federal Housing Finance Agency (FHFA) from the first quarter of 1991 to the first quarter of 2015. Quarterly state hpi data was obtained from `http://www.fhfa.gov/DataTools/Downloads/Documents/HPI/HPI_EXP_state.txt`. Figure 6.4 plots the hpi time series. We observe that there are two possible break points — one around 2006-2008 where hpi starts to depreciate after reaching a peak and one later around 2012 where hpi starts its revival. However, this merely suggests possible change points in terms of the overall mean level for hpi. Our focus here is the relation between house price and other economic factors. We want to see if there is a change point in time such that the model before and after the change point is different.

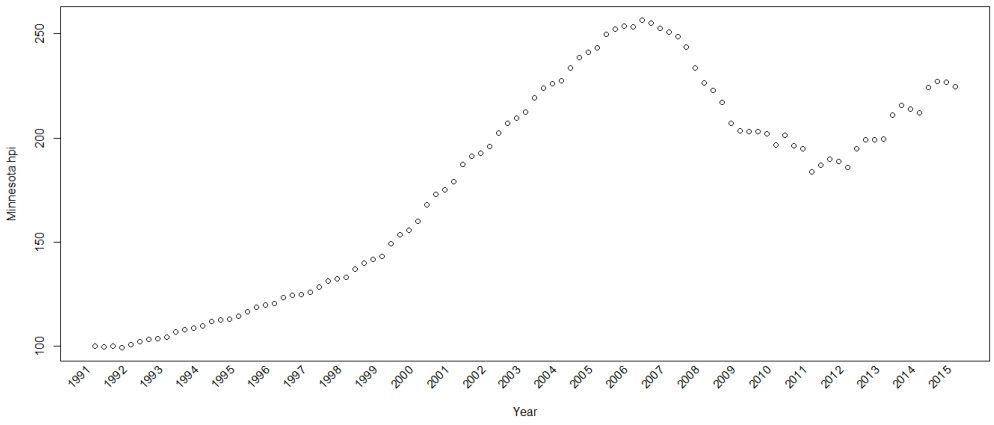


Figure 6.4: Minnesota hpi time series

The macro-economic indices used as explanatory variables include national unemployment rate (unemp) and national consumer price indices (cpi). The monthly cpi data was obtained from `http://inflationdata.com/inflation/consumer_price_index/`

historicalcpi.aspx?reloaded=true#Table?reloaded=true while the unemployment data was obtained from http://data.bls.gov/timeseries/LNS14000000. All monthly indices were averaged to convert to quarterly indices. Instead of including a national stock index in the model like the S&P 500 or the Dow Jones Industrial Average, we use the stock prices of Minnesota based Fortune 500 companies. 14 out of the 18 Minnesota-based Fortune 500 companies have been publicly traded since before 1991 and we include their stock prices in the regression model. Additionally, the list of top 10 employers in Minnesota available at http://mn.gov/deed/business/locating-minnesota/companies-employers/top-employers.jsp includes Wal-Mart Stores Inc. and Wells Fargo Bank Minnesota. Hence, the stock prices of these two companies are also included in the model. The 16 stocks used in total are listed in Table 6.6.

Table 6.6: List of stocks used in Minnesota hpi analysis

Company Name	Ticker Symbol	Company Name	Ticker Symbol
3M Company	MMM	St. Jude Medical, Inc.	STJ
Best Buy Co., Inc.	BBY	SuperValu, Inc.	SVU
Ecolab, Inc.	ECL	Target Corporation	TGT
Fastenal Co.	FAST	UnitedHealth Group Inc.	UNH
General Mills, Inc.	GIS	U.S. Bancorp	USB
Hormel Foods Corporation	HRL	Wal-Mart Stores, Inc.	WMT
Medtronic Plc.	MDT	Wells Fargo & Company	WFC
Mosaic Company	MOS	Xcel Energy Inc.	XEL

Financial indices often exhibit strong autocorrelation and consequently autoregressive components commonly feature in house price models (Nagaraja et al., 2011). Figure 6.5 plots the partial auto-correlation values of the hpi time series as a function of the lag. We observe that the index lagging one quarter behind (AR(1)) has very high correlation with the hpi time series but it quickly falls off beyond the first lag and all the subsequent lagged indices have insignificant partial correlations. Consequently, we include only the AR(1) term in the regression model.

Statistical analysis involving financial time series is often preceded by customary seasonality adjustment of the indices using standard time series techniques. It is well known that house price time series reveal a predictable and repetitive pattern with

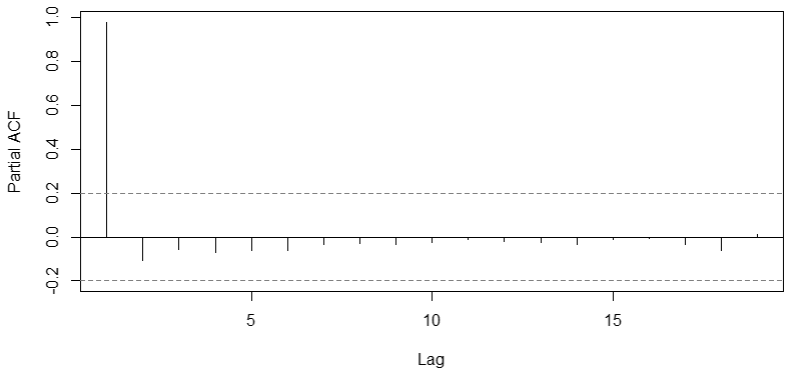


Figure 6.5: Partial autocorrelation function for Minnesota hpi time series

systematic highs in summer and lows in winter (Ngai and Tenreyro, 2014). Consequently, publishers of popular house price indices like the FHFA or Standard and Poor's (Case-Shiller index) produce a version of their indices discounting this effect (Federal Housing Finance Agency, 2014). However, Minnesota is a land of extreme climate experiencing one of the widest range of temperatures in U.S. It is of interest to investigate if the impact of weather in Minnesota on its house prices extends beyond the routine pattern. Hence, we include the state level quarterly average temperatures (temp) and precipitation (precip) in the model. The data is obtained from `http://www.ncdc.noaa.gov/cag/time-series`.

Our model is stated as follows. Assuming K-change points $\tau_1 < \tau_2 < \ldots < \tau_K$ where K is to de determined by the data, the regression model at the t^{th} quarter, for $\tau_{k-1} \leq t \leq \tau_k$ is given by:

$$\begin{aligned} hpi_t = \beta_k^{intercept} + \beta_k^{ar(1)}\text{hpi}_{t-1} + \beta_k^{cpi}\text{cpi}_t + \beta_k^{unemp}\text{unemp}_t \\ + \beta_k^{temp}\text{temp}_t + \beta_k^{precip}\text{precip}_t + \beta_k^{'stocks}\text{stocks}_t + \epsilon_t \end{aligned} \tag{6.7}$$

Here stocks_t denote the 16×1 vector formed by stacking up the stock prices at time t of the companies listed in Table 6.6 and β_k^{stocks} is the corresponding coefficient vector.

6.4.3 Results

We used the data from the second quarter of 1991 to the second quarter to 2014 for model fitting. The first quarter data of 1991 was used for the AR(1) term, whereas the data for last two quarters of 2014 and first quarter of 2015 were held out for out-of-sample validation. Under the assumption of K change points, separate regression models are fit to each of the $K+1$ segments. Hence, although the sample size ($n = 93$) was larger than the number of predictors ($p = 22$) in model 6.7, depending on the location of change-points, many segments may have less than 22 datapoints thereby necessitating high-dimensional regression methods. For example, a possible change point before before 1996 or after 2009 would imply that one of the data segments will have less than 22 observations and a classical least squares analysis cannot be conducted. Higher values of K (≥ 3) imply that average sample size for each segment ($n/(K+1)$ becomes really small and the estimates obtained may not be reliable. Hence, we restrict ourselves to the choices $K = 0, 1$ and 2 and fit model 6.7 using the BASAD priors for the coefficient vectors in each segment. Note that for $K = 0$, i.e., no change point, the model simply reduces to the traditional BASAD model. The models for different values of K were assessed based on their in-sample DIC score and out-of-sample RMSPE score (Yeniay and Goktas, 2002). Due to the presence of the autoregressive term, out-of-sample forecasts were obtained using one-step-ahead predictions.

The models for $K = 0, 1$ and 2 are denoted by $basad0$, $basad1$ and $basad2$ respectively. For comparison, we used the high-dimensional change-point lasso (Lee et al., to appear) where for each fixed value of τ_1, coefficients on either side were estimated using lasso. The optimal τ_1 is one which minimizes the mean squared predictive error. We refer to this model as $lasso1$. Additionally, to elucidate why low dimensional analysis is not suitable for this data, we also used two low dimensional models — a low dimensional one change point linear model ($lm1$) which is similar to $lasso1$ but uses classical least squares to estimate the coefficients for each τ_1, and a low dimensional Bayesian linear model $bayeslm1$ with one change point (similar to Carlin et al., 1992) with normal Inverse gamma (NIG) priors for $(\beta_1, \beta_2, \sigma^2)$ and uniform prior for τ_1.

Table 6.7 contains the DIC scores (only for the Bayesian models), RMSPE values and estimated change points for all the models. Both the DIC score and the RMSPE score for $K = 0$ were significantly worse than the scores for $K = 1$ and 2 justifying

Table 6.7: Minnesota hpi analysis: DIC, RMSPE scores and estimated change points

Model	DIC	RMSPE	$\hat{\tau}_1$	$\hat{\tau}_2$
basad0	287	4.72		
basad1	250	2.17	2008Q4 (2008Q2, 2009Q1)	
basad2	269	2.23	2006Q2 (2005Q4, 2007Q3)	2011Q1 (2010Q4, 2011Q2)
lasso1		2.71	2008Q2	
bayeslm1	288	10.00	2008Q3 (2008Q2, 2008Q4)	
lm1		23.31	2008Q2	

the use of a change point model. The single change point model detected a change point around late 2008- early 2009 which coincides with the sub-prime mortgage crisis. The two change point model detected change points in mid 2006 and early 2011. The DIC score for the single change point model was substantially better. RMSPE scores for change point models for time series data only validate the accuracy of the models after the last change point. We observed that the RMSPE score were similar for $K = 1$ and 2 with the former turning out to be marginally better. The change point lasso also estimated a change point around mid 2008 however the RMSPE score was higher than our Bayesian model. The low dimensional models $lm1$ and $bayeslm1$ were also able to detect a change point in 2008. However, their model evaluation metrics were significantly worse. This is not surprising as a change point in mid 2008 leaves less than 25 observations to estimate a 22-dimensional vector β_2. In a low dimension approach like linear least squares and, to a lesser extent, in Bayesian linear model, this will lead least to unstable estimates. This elucidates that in-spite of $n = 93$ being sufficiently larger than $p = 22$, in presence of a change point, the location of the change point may warrant a regularized approach to ensure numerically stable analysis.

We present the subsequent analysis only for the single change point model $basad1$ as both in-sample and out-of-sample validations provide strongest evidence in favor of a single change point. Figure 6.6 plots the probability of selection for each of the regressors in model (6.7) before and after the change point using BASAD priors. We observe that the set of variables selected by the median probability model differs before and after the change point. The AR(1) index and precipitation are selected with high probabilities in both segments. However, the selection of stocks differ considerably

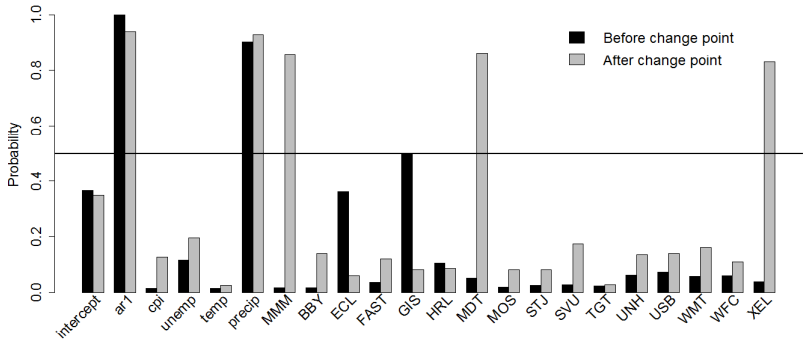


Figure 6.6: Minnesota hpi analysis: Posterior median probabilities of variable selection using single change point model

on either side of the change point. We see that prior to change point in 2008, there was little correlation between hpi and stocks with only General Mills (GIS) having a posterior median probability close to 0.5 (0.498). Perhaps this is a reflection of the fact discussed earlier that stock prices and hpi did not exhibit co-movements during the early 2000s. After the change point in 2008 the stocks of 3M (MMM), Medtronic (MDT) and Xcel Energy (XEL) are selected with high probability.

Turning to the actual coefficient values presented in Table 6.8, we observe that the value for the coefficient corresponding to the AR(1) index drops significantly post change point indicating less autoregressive behavior after the change point. We also observe a positive association of hpi with precipitation. Since summer months witness significantly higher precipitation than winter (see Figure 6.7), this merely corroborates

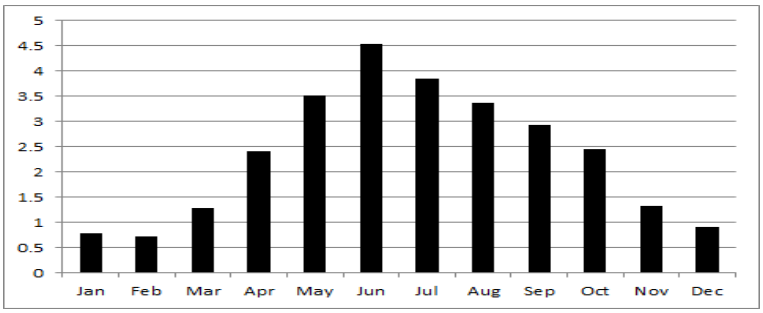


Figure 6.7: Avg. monthly precipitation in Minnesota between 1991Q1 to 2015Q1

Table 6.8: Minnesota hpi analysis: Posterior median (and 95% confidence intervals) for the coefficients. The variables which are selected in at least one segment are indicated by *.

	Before change point	After change point
intercept	-0.015 (-4.537, 2.65)	0.004 (-3.557, 4.003)
AR(1)*	1.034 (0.96, 1.106)	0.761 (0.022, 0.873)
cpi	0 (-0.055, 0.054)	0.042 (-0.07, 0.861)
unemp	-0.015 (-0.697, 0.137)	0.01 (-1.063, 1.304)
temp	-0.02 (-0.064, 0.044)	0.063 (-0.021, 0.121)
precip*	0.964 (-0.003, 1.63)	1.952 (-0.012, 2.806)
MMM*	0.019 (-0.065, 0.107)	-0.573 (-0.832, 0.015)
BBY	-0.003 (-0.091, 0.088)	0.065 (-0.071, 0.322)
ECL	-0.104 (-0.541, 0.048)	0.026 (-0.09, 0.217)
FAST	0 (-0.132, 0.132)	-0.033 (-0.746, 0.093)
GIS*	-0.151 (-0.958, 0.068)	-0.002 (-0.206, 0.185)
HRL	-0.013 (-0.636, 0.131)	0.016 (-0.12, 0.438)
MDT*	0.071 (-0.022, 0.168)	1.214 (-0.073, 1.666)
MOS	-0.056 (-0.107, -0.004)	-0.017 (-0.413, 0.074)
STJ	0.007 (-0.097, 0.107)	0.026 (-0.101, 0.394)
SVU	-0.035 (-0.144, 0.07)	0.021 (-0.129, 1.569)
TGT	0.001 (-0.108, 0.11)	0 (-0.116, 0.117)
UNH	-0.061 (-0.185, 0.043)	-0.042 (-0.521, 0.083)
USB	0.045 (-0.071, 0.241)	-0.017 (-2.98, 0.147)
WMT	0.073 (-0.016, 0.17)	0.014 (-0.122, 1.083)
WFC	-0.006 (-0.154, 0.14)	0.001 (-0.142, 2.296)
XEL*	-0.005 (-0.13, 0.112)	1.279 (-0.05, 2.436)

the traditional 'hot season cold season' trend of house prices. What is more interesting is the fact that this effect is much more pronounced after 2008 indicating more disparity between summer and winter house prices in the post-recession market.

The posterior predictive distributions for the hpi at each quarter was also obtained from the MCMC sampler and Figure 6.8 plots the true hpi versus the in-sample median

fits and confidence intervals. We see that that the posterior confidence intervals provide

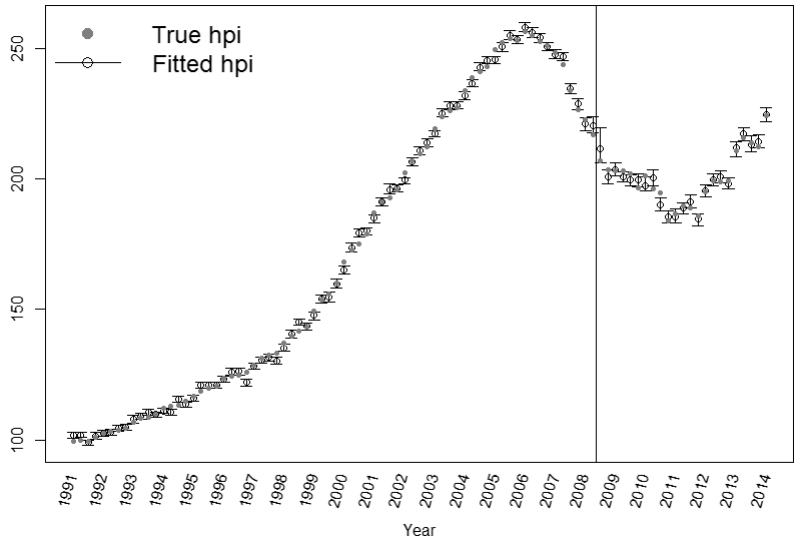


Figure 6.8: Minnesota hpi analysis: In-sample posterior predictive medians and confidence intervals of hpi. The vertical line indicates the posterior median estimate of the change point.

substantial coverage. We also observe that the confidence intervals are wider after the change point. This is expected as there are only about 23 time points compared to around 70 before the change point. The out-of-sample posterior fits are provided in Figure 6.9. All the three out-of-sample predictions are very close to their true values. However, the out-of-sample confidence intervals are significantly wider reflecting the uncertainty associated with prediction.

Observe from Figure 6.6 that in presence of the stock prices of Minnesota based companies, national level macro-economic indicators like the cpi or unemployment were not selected in the model. This perhaps provides evidence in support of the conjecture that hpi is strongly correlated with local macro-economics (Garmaise and Moskowitz, 2002). However, its worthwhile to point out that multivariate regression models, although a simple and powerful tool to determine correlation, rarely imply causality. Any confirmatory assessment of the change points detected and the variables selected by our method would require further economic research. Nevertheless, the model evaluation

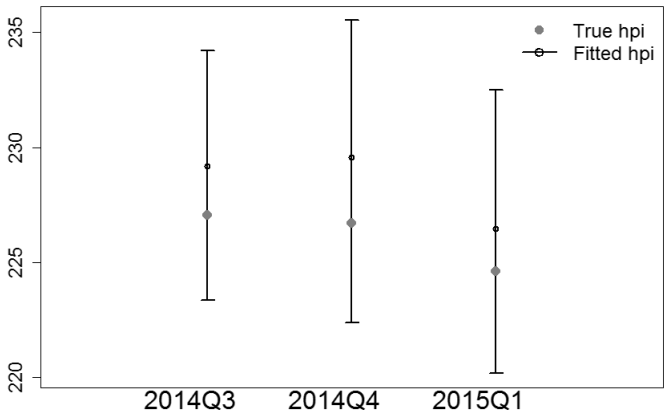


Figure 6.9: Minnesota hpi analysis: Out-of-sample posterior predictive medians and confidence intervals of hpi.

metrics in Table 6.7 provide very strong evidence in favor of one or more change points thereby justifying the use of our methodology to analyze the data.

6.5 Conclusion

We have presented a very general method for analyzing high dimensional data using changing linear regression. Our fully Bayesian approach offers several inferential advantages including quantifying uncertainty regarding the change points as well as variable selection for each segment. Our framework is flexible to the choice of variable selection priors although the BASAD prior empirically outperformed other competing choices. A wide range of constrained variable selections like grouping or partial selection can be seamlessly accomplished in our setup. The analysis of Minnesota hpi data using our method revealed strong evidence for a potential change point with respect to the association with other macro-economic variables.

We have discussed several approaches for handling unknown number of change points. However, most of them comes with warnings regarding computational requirements. More efficient algorithms for simultaneous detection of number of change points need to be researched. Other potential extensions include accommodating missing data,

measurement errors or non-Gaussian responses in a high dimensional changing regression setup. Extensions to change point detection in high dimensional VAR models also need to be explored due to the extensive usage of VAR models in economics research (Christiano et al., 2000; Bernanke et al., 2004). In a time series context, our work is restricted to detecting historical change points. Detecting future change points in high dimensional time series is equally important to provide accurate predictions. We identify all these areas as directions for future research.

References

Adams, R. P. and MacKay, D. J. (2007), "Bayesian Online Changepoint Detection," *arXiv preprint arXiv:0710.3742*.

Ahamada, I. and Diaz Sanchez, J. L. (2013), "A Retrospective Analysis of the House Prices Macro-relationship in the United States," *The World Bank: Policy Research Working Paper Series*, Jul.

Allcroft, D. J. and Glasbey, C. A. (2003), "A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation," *Journal of the Royal Statistical Society, Series C*, 52, 487–498.

Apergis, N. (2003), "Housing Prices and Macroeconomic Factors: Prospects within the European Monetary Union," *International Real Estate Review*, 6, 63–47.

Assuncao, R. and Krainski, E. (2009), "Neighborhood dependence in Bayesian spatial models," *Biometrical Journal*, 51, 851–869.

Bai, Y., Song, P. X. K., and Raghunathan, T. E. (2012), "Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes," *Journal of the Royal Statistical Society, Series B*, 74, 799–824.

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., and Morris, J. S. (2010), "Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data," *Journal of the American Statistical Association*, 105, 1358–1375.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC, 2nd ed.

Banerjee, S., Finley, A. O., Waldmann, P., and Ericcson, T. (2010), "Hierarchical Spatial Process Models for Multiple Traits in Large Genetic Trials," *Journal of the American Statistical Association*, 105, 506–521.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Datasets," *Journal of the Royal Statistical Society, Series B*, 70, 825–848.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal Predictive Model Selection," *Annals of Statistics*, 32, 870–897.

Barry, D. and Hartigan, J. A. (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*, 88, 309–319.

Bechtold, W. A. and Patterson, P. L. (2005), "The Enhanced Forest Inventory and Analysis National Sample Design and Estimation Procedures," *SRS-80, U.S. Department of Agriculture, Forest Service, Southern Research Station: Asheville, NC.*

Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2014a), "An $\ell_1, \ell_2, \ell_\infty$-Regularization Approach to High-Dimensional Errors-in-variables Models," *http://arxiv.org/abs/1412.7216.*

— (2014b), "Linear and Conic Programming Estimators in High-Dimensional Errors-in-variables Models," *http://arxiv.org/abs/1408.0241.*

Benjamini, Y. and Speed, T. (2012), "Estimation and correction for GC-content bias in high throughput sequencing," *Nucleic Acids Research*, 40, 72.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2004), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," Working Paper 10220, National Bureau of Economic Research.

Besag, J. (1974), "Spatial interaction and statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–225.

Bevilacqua, M., Fass, ò. A., Gaetan, C., and Velandia, D. (2015), "Covariance tapering for multivariate Gaussian random fields estimation," *Statistical Methods and Application.*

Bevilacqua, M. and Gaetan, C. (2014), "Comparing Composite Likelihood Methods Based on Pairs for Spatial Gaussian Random Fields," *Statistics and Computing*, 1–16.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012), "Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach," *Journal of the American Statistical Association*, 107, 268–280.

Bickel, P. J. and Levina, E. (2008a), "Covariance regularization by thresholding," *The Annals of Statistics*, 36, 2577–2604.

— (2008b), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36, 199–227.

Birmili, W., Schepanski, K., Ansmann, A., Spindler, G., Tegen, I., Wehner, B., Nowak, A., Reimer, E., Mattis, I., Muller, K., Bruggemann, E., Gnauk, T., Herrmann, H., Wiedensohler, A., Althausen, D., Schladitz, A., Tuch, T., and Loschau, G. (2008), "A Case of Extreme Particulate Matter Concentrations over Central Europe Caused by Dust Emitted over the Southern Ukraine," *Atmospheric Chemistry and Physics*, 8, 997–1016.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Van Dingenen, R., van Donkelaar, A., and Thurston, G. D. (2011), "Exposure Assessment for Estimation of the Global Burden of Disease Attributable to Outdoor Air Pollution," *Environmental Science and Technology*, 46, 652–660.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003), "Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions," *Journal of the Royal Statistical Society: Series B*, 65, 3–39.

Brunekreef, B. and Holgate, S. T. (2002), "Air pollution and health," *The Lancet*, 360, 1233–1242.

Buhlmann, P. and van de Geer, S. A. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York, NY: Springer.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010), "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, 38, 2118–2144.

Candès, E. and Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, 35, 2313–2351.

Candiani, G., Carnevale, C., Finzi, G., Pisoni, E., and Volta, M. (2013), "A comparison of reanalysis techniques: Applying optimal interpolation and Ensemble Kalman Filtering to improve air quality monitoring at mesoscale," *Science of the Total Environment*, 458-460, 7–14.

Carlin, B. P. and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society: Series B*, 57, 473–484.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992), "Hierarchical Bayesian Analysis of Changepoint Problems," *Journal of the Royal Statistical Society: Series C*, 41, 389–405.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480.

Case, K. E., Shiller, R. J., and Quigley, J. M. (2001), "Comparing Wealth Effects: The Stock Market Versus the Housing Market," Working Paper 8606, National Bureau of Economic Research.

Catte, P., Girouard, N., Price, R., and Andre, C. (2004), "Housing Markets, Wealth and the Business Cycle." OECD Working Paper, No 394.

Chen, J. and Gupta, A. K. (1997), "Testing and Locating Variance Changepoints with Application to Stock Prices," *Journal of the American Statistical Association*, 92, 739–747.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2000), "Monetary policy shocks: What have we learned and to what end?" in *Handbook of Macroeconomics*, eds. Taylor, J. and Woodford, M., Amsterdam, Holland: Elsevier, pp. 65 – 148.

Clayton, D. G. and Bernardinelli, L. (1992), "Bayesian Methods for Mapping Disease Risk," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. Elliott, P., Cuzick, J., English, D., and Stern, R., Oxford University Press, pp. 205–220.

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008), "Bivariate Binomial Spatial Modeling of Loa Loa Prevalence in Tropical Africa," *Journal of the American Statistical Association*, 103, 21–37.

Cressie, N. and Huang, H. (1999), "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of American Statistical Association*, 94, 1330–1340.

Cressie, N. A. C. and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Data Sets," *Journal of the Royal Statistical Society, Series B*, 70, 209–226.

Cressie, N. A. C., Shi, T., and Kang, E. L. (2010), "Fixed rank filtering for spatio-temporal data," *Journal of Computational and Graphical Statistics*, 19, 724–745.

Cressie, N. A. C. and Wikle, C. K. (2011), *Statistics for spatio-temporal data*, Hoboken, NJ: Wiley, Wiley Series in Probability and Statistics.

Dagum, L. and Menon, R. (1998), "OpenMP: An Industry Standard API for Shared-memory Programming," *Computational Science & Engineering, IEEE*, 5, 46–55.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association (In press)*.

Davis, T. A. (2006), *Direct Methods for Sparse Linear Systems*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing*, 12, 27–36.

Denby, B., Schaap, M., Segers, A., Builtjes, P., and Horalek, J. (2008), "Comparison of Two Data Assimilation Methods for Assessing PM10 Exceedances on the European Scale," *Atmospheric Environment*, 42, 7122–7134.

Denby, B., Sundvor, I., Cassiani, M., de Smet, P., de Leeuw, F., and Horalek, J. (2010), "Spatial Mapping of Ozone and SO2 Trends in Europe," *Science Of The Total Environment*, 408, 4795–4806.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based Geostatistics (with discussion)," *Applied Statistics*, 47, 299–350.

Du, J., Zhang, H., and Mandrekar, V. S. (2009), "Fixed-domain Asymptotic Properties of Tapered Maximum Likelihood Estimators," *Annals of Statistics*, 37, 3330–3361.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008), "Efficient Projections Onto the L1-ball for Learning in High Dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, New York, NY, USA: ACM, ICML '08, pp. 272–279.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

Efron, B., Hastie, T., and Tibshirani, R. (2007), "Discussion: The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," *The Annals of Statistics*, 35, 2358–2364.

Ehlers, R. S. and Brooks, S. P. (2008), "Adaptive Proposal Construction for Reversible Jump MCMC," *Scandinavian Journal of Statistics*, 35, 677–690.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models with Block Composite Likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315.

El Karoui, N. (2008), "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *The Annals of Statistics*, 36, 2717–2756.

Emory, X. (2009), "The Kriging Update Equations and Their Application to the Selection of Neighboring Data," *Computational Geosciences*, 13, 269–280.

European Commission (2015), "European Union Air Quality Standards," *http://ec.europa.eu/environment/air/quality/standards.htm*.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

— (2006), *Statistical challenges with high dimensionality: Feature selection in knowledge discovery*, Eur. Math. Soc., Zürich, p. 595–622.

Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, 20, 101–148.

Fan, Y. and Sisson, S. (2011), "Reversible Jump Markov chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo*, eds. Brooks, S. P., Gelman, A., Jones, G. L., and Meng, X.-L., Boca Raton, FL: Chapman & Hall/CRC, pp. 67–87.

Farr, W. M., Mandel, I., and Stevens, D. (2015), "An Efficient Interpolation Technique for Jump Proposals in Reversible-jump Markov Chain Monte Carlo Calculations," *Royal Society Open Science*, 2.

Federal Housing Finance Agency (2014), "U.S. House Price Index Report - 3Q 2014 / September 2014," Technical report.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2012), "Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes," *Journal of Geographical Systems*, 14, 29–47.

— (2013), "spBayes for Large Univariate and Multivariate Point-referenced Spatio-temporal Data Models," *Journal of Statistical Software*, 0, In press.

Finley, A. O., Banerjee, S., and McRoberts, R. E. (2009), "Hierarchical Spatial Models for Predicting Tree Species Assemblages across Large Domains," *The Annals of Applied Statistics*, 0, 1–32.

Flemming, J., Inness, A., Flentje, H., Huijnen, V., Moinat, P., Schultz, M. G., and Stein, O. (2009), "Coupling global chemistry transport models to ECMWF's integrated forecast system," *Geoscientific Model Development*, 2, 253–265.

Friedman, J., Hastie, T., and Tibshirani, R. (2007), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Furrer, R., Genton, M. G., and Nychka, D. W. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 503–523.

Garmaise, M. J. and Moskowitz, T. J. (2002), "Confronting Information Asymmetries: Evidence from Real Estate Markets," Working Paper 8877, National Bureau of Economic Research.

Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P. (2010), *Handbook of Spatial Statistics*, Boca Raton, FL: CRC Press.

Gelfand, A. E. and Banerjee, S. (2010), "Multivariate Spatial Process Models," in *Handbook of Spatial Statistics*, eds. Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., Boca Raton, FL: Chapman & Hall/CRC, pp. 495–516.

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005a), "Spatial Process Modelling for Univariate and Multivariate Dynamic Spatial Data," *Environmetrics*, 16, 465–479.

— (2005b), "Spatial process modelling for univariate and multivariate dynamic spatial data," *Environmetrics*, 16, 465–479.

Gelfand, A. E. and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial Modeling with Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Gneiting, T. (2002), "Nonseparable, stationary covariance functions for spacetime data," *Journal of American Statistical Association*, 97, 590–600.

Gneiting, T., Genton, M. G., and Guttorp, P. (2007), "Geostatistical space-time models, stationarity, separability and full symmetry," in *Statistics of SpatioTemporal Systems*, Boca Raton, FL: Chapman & Hall/CRC, pp. 151–175, (eds Finkenstaedt, B. and Held, L. and Isham, V.).

Gneiting, T. and Guttorp, P. (2010), "Continuous-parameter Spatio-temporal Processes," in *Handbook of Spatial Statistics*, eds. Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., Boca Raton, FL: Chapman & Hall/CRC, pp. 427–436.

Gräler, B., Gerharz, L., and Pebesma, E. (2011), "Spatio-temporal Analysis and Interpolation of PM10 Measurements in Europe," *ETC/ACM Technical Paper*, 10.

Gramacy, R. B. and Apley, D. W. (2014), "Local Gaussian Process Approximation for Large Computer Experiments," *http://arxiv.org/abs/1303.0383*.

Gramacy, R. B. and Lee, H. (2008), "Bayesian Treed Gaussian Process Models with an Application to Computer Experiments," *Journal of the American Statistical Association*, 103, 1119–1130.

Gramacy, R. B., Niemi, J., and Weiss, R. M. (2014), "Massively Parallel Approximate Gaussian Process Regression," *http://arxiv.org/abs/1310.5182*.

Green, P. J. (1995), "Reversible Jump Markov chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Green, P. J. and Hastie, D. I. (2009), "Reversible Jump MCMC," .

Hamm, N. A. S., Finley, A. O., Schaap, M., and Stein, A. (2015), "A Spatially Varying Coefficient Model for Mapping PM10 Air Quality at the European scale," *Atmospheric Environment*, 102, 393–405.

Han, C. and Carlin, B. P. (2001), "Markov Chain Monte Carlo Methods for Computing Bayes Factors," *Journal of the American Statistical Association*, 96, 1122–1132.

Hastie, D. I. and Green, P. J. (2012), "Model Choice Using Reversible Jump Markov Chain Monte Carlo," *Statistica Neerlandica*, 66, 309–338.

Hastie, T., Tibshirani, R., and Friedman, J. (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer, 2nd ed.

Hendriks, C., Kranenburg, R., Kuenen, J., van Gijlswijk, R., Kruit, R. W., Segers, A., van der Gon, H. D., and Schaap, M. (2013), "The origin of ambient particulate matter concentrations in the Netherlands," *Atmospheric Environment*, 69, 289–303.

Higdon, D. (2001), "Space and Space Time Modeling using Process Convolutions," *Technical Report, Institute of Statistics and Decision Sciences, Duke University, Durham.*

Hodges, S. J. (2013), *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*, Chapman and Hall, 1st ed.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. D. (2013), "Long-term Air Pollution Exposure and Cardio- respiratory Mortality: A Review," *Environmental Health*, 12, 43.

Hofmann, B. (2003), "Bank Lending and Property Prices: Some International Evidence." Working papers, Hong Kong Institute for Monetary Research.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance matrix selection and estimation via penalised normal likelihood," *Biometrika*, 93, 85–98.

Intel (2015), "Math Kernel Library," http://developer.intel.com/software/products/mkl/.

Ishwaran, H. and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773.

Jones, R. H. and Zhang, Y. (1997), "Models for continuous stationary space-time processes," in *Modelling Longitudinal and Spatially Correlated Data*, New York, NY: Springer, pp. 289–298, (eds T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger).

Kammann, E. E. and Wand, M. P. (2003), "Geoadditive Models," *Applied Statistics*, 52, 1–18.

Katzfuss, M. (2016), "A multi-resolution approximation for massive spatial datasets," *Journal of the American Statistical Association.*

Katzfuss, M. and Cressie, N. (2012), "Bayesian Hierarchical Spatio-temporal Smoothing for Very Large Datasets," *Environmetrics*, 23, 94–107.

Kaufman, C. G., Scheverish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555.

Kezim, B. and Pariseau, S. E. (2004), "Bayesian Analysis of a Structural Change in Volatility Using the Gibbs Sampler with an Application to Stock Market Returns," *Journal of Business and Economic Studies*, 10, 1–11.

Kyriakidis, P. C. and Journel, A. G. (1999), "Geostatistical space-time models: a review," *Mathematical Geology*, 31, 651–684.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation," *The Annals of Statistics*, 37, 4254–4278.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford, United Kingdom: Clarendon Press.

Lee, S., Seo, M. H., and Shin, Y. (to appear), "The lasso for high dimensional regression with a possible change point," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a–n/a.

Lenardon, M. J. and Amirdjanova, A. (2006), "Interaction Between Stock Indices via Changepoint Analysis," *Applied Stochastic Models in Business and Industry*, 22, 573–586.

Leroux, B. G., Lei, X., and Breslow, N. (2000), "Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, eds. Halloran, M. E. and Berry, D., New York, NY: Springer New York, pp. 179–191.

Leung, C. K. Y. (2004), "Macroeconomics and Housing: A Review of the Literature," Discussion Papers 00004, Chinese University of Hong Kong, Department of Economics.

Levina, E., Rothman, A., and Zhu, J. (2008), "Sparse estimation of large covariance matrices via a nested Lasso penalty," *Ann. Appl. Stat.*, 2, 245–263.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000), "Smoothing Spline ANOVA Models for Large Data Sets with Bernoulli Observations and the Randomized GACV," *Annals of Statistics*, 28, 1570–1600.

Littenberg, T. B. and Cornish, N. J. (2009), "Bayesian Approach to the Detection Problem in Gravitational Wave Astronomy," *Physical Review*, 80, 063007–1– 063007–19.

Lloyd, C. D. and Atkinson, P. M. (2004), "Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data," *International Journal of Applied Earth Observation and Geoinformation*, 5, 293–305.

Loh, P. and Wainwright, M. J. (2012), "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," *The Annals of Statistics*, 40, 1637–1664.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., and Straif, S. (2013), "The carcinogenicity of outdoor air pollution," *The Lancet Oncology*, 14, 1262–1263.

MacNab, Y. and Dean, C. (2000), "Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models," *Statistis in Medicine*, 19, 15–30.

Mai, Q., Zou, H., and Yuan, M. (2012), "A direct approach to sparse discriminant analysis in ultra-high dimensions," *Biometrika*, 99, 29–42.

Malpezzi, S. (2003), "Hedonic Pricing Models: A Selective and Applied Review," in *Housing Economics and Public Policy*, eds. OSullivan, T. and Gibb, K., Malder, MA: Blackwell Science Ltd, pp. 67–89.

Manders, A. M. M., Schaap, M., and Hoogerbrugge, R. (2009), "Testing the Capability of the Chemistry Transport Model LOTOS-EUROS to Forecast PM10 Levels in the Netherlands," *Atmospheric Environment*, 43, 4050–4059.

Mavrodiy, A. (2005), "Factor Analysis of Real Estate Prices," *EERC MA thesis*.

McCulloch, R. E. and Tsay, R. S. (1993), "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88, 968–978.

Meinshausen, N. and Buhlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462.

Mues, A., Kuenen, J., Hendriks, C., Manders, A., Segers, A., Scholz, Y., Hueglin, C., Builtjes, P., and Schaap, M. (2014), "Sensitivity of air pollution simulations with LOTOS-EUROS to the temporal distribution of anthropogenic emissions," *Atmospheric Chemistry and Physics*, 14, 939–955.

Nagaraja, C. H., Brown, L. D., and Zhao, L. H. (2011), "An Autoregressive Approach to House Price Modeling," *The Annals of Applied Statistics*, 5, 124–149.

Narisetty, N. N. and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817.

Ngai, L. R. and Tenreyro, S. (2014), "Hot and Cold Seasons in the Housing Market," *American Economic Review*, 104, 3991–4026.

Omidi, M. and Mohammadzadeh, M. (2015), "A New Method to Build Spatio-temporal Covariance Functions: Analysis of Ozone Data," *Statistical Papers*, 1–15.

Ong, T. S. and Chang, Y. S. (2013), "Macroeconomic Determinants of Malaysian Housing Market," *Journal of Human and Social Science Research*, 1, 119–127.

Otrok, C. and Terrones, M. E. (2005), "House Prices, Interest Rates and Macroeconomic Fluctuations: International Evidence," *Unpublished manuscript*.

Panagiotidis, T. and Printzis, P. (2015), "On the Macroeconomic Determinants of the Housing Market in Greece: a VECM approach," *LSE Research Online Documents on Economics*.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Pfeifer, P. E. and Deutsch, S. J. (1980a), "Independence and sphericity tests for the residuals of spacetime ARMA models," *Communications in Statistics - Simulation and Computation*, 9, 533–549.

— (1980b), "Stationarity and invertibility regions for low order STARMA models." *Communications in Statistics - Simulation and Computation*, 9, 551–562.

Pouliot, G., Pierce, T., van der Gon, H. D., Schaap, M., Moran, M., and Nopmongcol, U. (2012), "Comparing emission inventories and model-ready emission datasets between Europe and North America for the AQMEII project," *Atmospheric Environment*, 53, 4–14.

Purdom, E. and Holmes, S. P. (2005), "Error Distribution for Gene Expression Data," *Statistical Applications in Genetics and Molecular Biology*, 4.

Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009), "The Bayesian Group-Lasso for Analyzing Contingency Tables," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA: ACM, ICML '09, pp. 881–888.

Rasmussen, C. E. and Williams, C. K. I. (2005), *Gaussian Processes for Machine Learning*, Cambridge, MA: The MIT Press, 1st ed.

Reeves, J., Chen, J., Wang, X., Lund, R., and Lu, Q. (2007), "A Review and Comparison of Changepoint Detection Techniques for Climate Data," *Journal of Applied Meteorology and Climatology*, 46, 900–915.

Renigier-Biozor, M. and Winiewski, R. (2013), "The Impact of Macroeconomic Factors on Residential Property Prices Indices in Europe," *Aestimum*, 0, 149–166.

R'Honi, Y., Clarisse, L., Clerbaux, C., Hurtmans, D., Duflot, V., Turquety, S., Ngadi, Y., and Coheur, P. F. (2013), "Exceptional Emissions of NH3 and HCOOH in the 2010 Russian Wildfires," *Atmospheric Chemistry and Physics*, 13, 4171–4181.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)," *Journal of the Royal Statistical Society: Series B*, 59, 731–792.

Rosenbaum, M. and Tsybakov, A. B. (2010), "Sparse recovery under matrix uncertainty," *Annals of Statistics*, 38, 2620–2651.

— (2013), "Improved matrix uncertainty selector," *IMS Collections. From probability to statistics and back: High dimensional models and processes*, 9, 276–290.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, 2, 494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields : Theory and Applications*, Boca Raton, FL: Chapman & Hall/CRC, Monographs on Statistics and Applied Probability.

Sang, H. and Huang, J. Z. (2012), "A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets," *Journal of the Royal Statistical Society, Series B*, 74, 111–132.

Schaap, M., Timmermans, R. M. A., Roemer, M., Boersen, G. A. C., Builtjes, P., Sauter, F., Velders, G., and Beck, J. (2008), "The LOTOS-EUROS model: description, validation and latest developments," *International Journal of Environment and Pollution*, 32, 270–290.

Schabenberger, O. and Gotway, C. A. (2004), *Statistical Methods for Spatial Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC, 1st ed.

Shaby, B. A. (2012), "The Open-faced Sandwich Adjustment for MCMC using Estimating Functions," *http://arxiv.org/abs/1204.3687*.

Shaby, B. A. and Ruppert, D. (2012), "Tapered Covariance: Bayesian Estimation and Asymptotics," *Journal of Computational and Graphical Statistics*, 21, 433–452.

Slijepcevic, S., Megerian, S., and Potkonjak, M. (2002), "Location Errors in Wireless Embedded Sensor Networks: Sources, Models, and Effects on Applications," *Mobile Computing and Communications Review*, 6, 67–78.

Sørensen, Ø., Frigessi, A., and Thoresen, M. (2013), "Measurement Error in LASSO: Impact and Likelihood Bias Correction," *Statistica sinica*, 23, Preprint.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York, NY: Springer, 1st ed.

— (2002), "The Screening Effect in Kriging," *The Annals of Statistics*, 30, 298–323.

— (2005), "Spacetime covariance functions," *Journal of American Statistical Association*, 100, 310–321.

— (2007), "Spatial Variation of Total Column Ozone on a Global Scale," *Annals of Applied Statistics*, 1, 191–210.

— (2008), "A Modeling Approach for Large Spatial Datasets," *Journal of the Korean Statistical Society*, 37, 3–10.

— (2013), "On a Class of Spacetime Intrinsic Random Functions," *Bernoulli*, 19, 387–408.

— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19.

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society, Series B*, 66, 275–296.

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., and Kerschbaumer, A. (2008), "A model inter-comparison study focussing on episodes with elevated PM10 concentrations," *Atmospheric Environment*, 42, 4567–4588.

Stoffer, D. S. (1986), "Estimation and identification of spacetime ARMAX models in the presence of missing data," *Journal of the American Statistical Association*, 81, 762–772.

Strömberg, P., Hedman, M., and Broberg, M. (2011), "Forecasting the House Price Index in Stockholm County 2011-2014: A Multiple Regression Analysis of Four Influential Macroeconomic Variables," *Bachelor's thesis from Mlardalen University / School of Sustainable Development of Society and Technology.*

Stroud, J. R., Muller, P., and Sanso, B. (2001), "Dynamic models for spatiotemporal data," *Journal of the Royal Statistical Society, Series B*, 63, 673–689.

Stroud, J. R., Stein, M. L., and Lysen, S. (2014), "Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice," *http://arxiv.org/abs/1402.4281.*

Tibshirani, R. (1994), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society Series B*, 91–108.

Tsatsaronis, K. and Zhu, H. (2004), "What Drives Housing Price Dynamics: Cross-country Evidence," *BIS Quarterly Review*, March, 65–78.

Turner, R., Saatci, Y., and Rasmussen, C. (2009), "Adaptive Sequential Bayesian Change Point Detection," .

van de Geer, S. A. and Buhlmann, P. (2009), "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

van de Kassteele, J. and Stein, A. (2006), "A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output," *Environmetrics*, 17, 309–322.

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society, Series B*, 50, 297–312.

— (1992), "A New Method of Prediction for Spatial Regression Models with Correlated Errors," *Journal of the Royal Statistical Society, Series B*, 54, 813–830.

Vershynin, R. (2011), "Introduction to the non-asymptotic analysis of random matrices. Compressed sensing," *arXiv:1011.3027*, 210–268.

Wagaman, A. S. and Levina, E. (2009), "Discovering Sparse Covariance Structures With the Isomap," *Journal of Computational and Graphical Statistics*, 18, 551–572.

Wainwright, M. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using l_1-Constrained Quadratic Programming (Lasso)," *IEEE transactions on information theory*, 55, 2183–2202.

Wall, M. (2004), "A close look at the spatial structure implied by the CAR and SAR models," *Journal of Statistical Planning and Inference*, 121, 311–324.

Wang, Q., Adiku, S., Tenhunen, J., et al. (2005), "On the Relationship of NDVI with Leaf Area Index in a Deciduous Forest Site," *Remote Sensing of Environment*, 94, 244–255.

Whittle, P. (1954), "On Stationary Processes in the Plane," *Biometrika*, 41, 434–449.

Wikle, C. and Cressie, N. A. C. (1999), "A Dimension-reduced Approach to Space-time Kalman Fltering," *Biometrika*, 86, 815–829.

Wu, W. and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika*, 90, 831–844.

Xu, G., Liang, F., and Genton, M. G. (2014), "A Bayesian Spatio-Temporal Geostatistical Model with an Auxiliary Lattice for Large Datasets," *Statistica Sinica*, In press.

Xue, L., Ma, S., and Zou, H. (2012), "Positive-Definite ℓ_1-Penalized Estimation of Large Covariance Matrices," *Journal of the American Statistical Association*, 107, 1480–1491.

Yeniay, O. and Goktas, A. (2002), "A Comparison of Partial Least Squares Regression with Other Prediction Methods," *Hacettepe Journal of Mathematics and Statistics*, 31, 99–111.

Zhang, X. and Kondraguanta, S. (2006), "Estimating Forest Biomass in the USA Using Generalized Allometric Models and MODIS Land Products," *Geophysical Research Letters*, 33, L09402.

Zhao, P. and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.